Historical trades, skills and agglomeration economies

Philipp Ehrl and Leonardo Monteiro Monasterio

Universidade Católica de Brasília, IPEA

7 April 2016

Historical trades, skills and agglomeration economies*

Philipp Ehrl‡
Universidade Católica de Brasília

Leonardo Monasterio
Universidade Católica de Brasília, IPEA, UCLA

April 7, 2016

Abstract

We exploit differences in the spatial distribution of industrial and liberal occupations in the years 1872 and 1920 to instrument for today's concentration of interpersonal and analytical skills in Brazil. The data suggest that the local supply of knowledge and manufacturing provided by these historical trades favored a growth path that has shaped the occupational structure until the present day, whereby the existence of a large local consumer market was a necessary condition for this development. By means of these instruments, we present causal evidence that the regional concentration of interpersonal and analytical skills generates positive wage externalities. Particularly university graduates and workers without formal education benefit most from these agglomeration economies.

Keywords: agglomeration economies, skills, long-run industrial development, Brazil
JEL Classification: R12, J31, C26, N16

*We thank Daniel da Mata and seminar participants at the UCLA, NARSC, CIDE, IPEA in Brasília and the UCB for the discussions and valuable comments. Ehrl gratefully acknowledges financial support from CAPES.

‡corresponding author: Universidade Católica de Brasília, Post-Graduate Program in Economics, SGAN 916, 70790-160 - Brasília, Brazil. Phone: +55 (61) 3448-7136. Mail: philipp.ehrl@ucb.br

1 Introduction

> *"(2) The scale of an industrial plant depends on the demand for its products [...] (7) Since it takes machines to produce machines, and these are themselves the product of many different factories and workshops, machinery is produced efficiently only in a place where factories and workshops are close enough together to help each other work in unison, i.e. in large towns".*
>
> — (von Thünen 1826/1966: 287-90), *The Isolated State*

Plants benefit from locating close to each other in large towns for two reasons listed in the initial quote: the positive interaction between economies of scale (supply) and the consumer market (demand); and the benefits due to the cooperation and exchange between different producers. In consonance, Marshall (1890) used the famous expression "mysteries of the trades" to paraphrase how people benefit from a concentration of workers with the "same skilled trade" through positive externalities. Arrow (1962) provides the related insight that physical production itself is required for "learning by doing". We aim to test the empirical content of these arguments, i.e., we evaluate the linkages among the concentration of knowledge and production, agglomeration economies and long-run growth.

The present paper makes two main contributions to the literature. (1) As far as we know, we are the first to present causal evidence that the spatial concentration of interactive and analytic skills generates positive externalities for workers' wages.[1] A general problem with the identification of agglomeration economies is their endogeneity due to reverse causality and the sorting of workers to high-wage regions according to unobservable abilities (Combes *et al.* 2008). (2) To tackle this endogeneity problem and the measurement error in the skill variables, we exploit differences in the spatial distribution of industrial and liberal occupations from the Brazilian Censuses in 1872 and 1920 to construct instrumental variables. The identification assumption rests on the argument that the regional concentration of these professions in the past stimulated sustainable long-run economic growth and thus resulted in a high conglomeration of analytic and interactive skills today.

A preliminary analysis of the Brazilian data reveals that interactive and analytic skills are heavily concentrated in densely populated regions. The same applies to managers, scientists and skilled technicians, which correspond to the type of occupations with the highest wages/skills. The corollary is that manual skills can primarily be found in rural regions. Maciente (2013) also studies the distribution of skills in Brazil with different skill definitions but comes to the same basic conclusions. Similar observations are reported for the US by Bacolod *et al.* (2009b), Florida *et al.* (2012) and as well as for Sweden by Andersson *et al.* (2014). Moreover, Bacolod *et al.* (2009a) observe that returns to cognitive and interactive skills grow with the size of the region. These results are suggestive that in the modern economy, the agglomeration economies mainly arise from the interpersonal exchange of ideas and knowledge spillovers. Thus far, these insights have been merely

[1] Compare the survey on the empirics of agglomeration economies by Combes and Gobillon (2015).

based on simple correlations so to make progress instrumental variables are introduced here. We are convinced that the chosen instruments are valid for the following reasons.

Industrialist is one of the professions distinguished in the Census of 1872, the era of the very beginning of the industrialization in Brazil. Industrialists were the owners of factories and their factories provided jobs and generated capital which could be reinvested to expand and continually modernize the production. On the other hand, liberal professions (such as lawyers, judges, professors and teachers) generated knowledge – a scarce and highly valuable asset at that time – and contributed to the functioning of the fragile institutions. As argued above, the local accumulation of knowledge and physical production were key to productivity increases and the sustained economic growth that eventually led to the local development of ever more modern and dynamic industries. A region with predominantly agricultural workers, domestic servants and individuals without a profession, in contrast, hindered the development of the local economy and preserved the ancient structures. An econometric test of these theories over a long period is, to the best of our knowledge, absent thus far. Moreover, we observe that positive progress was much more likely if the region was a large agglomeration already. This suggests that the interaction between supply (provided by the industrialists) and consumer demand created a virtuous cycle for the local economy. Thus, market potential, circular linkages, and history do play a crucial role in the development of agglomerations, as predicted by the theories of the New Economic Geography (Fujita *et al.* 2001).

Since Ciccone and Hall (1996), regional economists have used historical variables to instrument for the population size or density of regions, exploiting the persistence of large settlements per se. Alternatively, Rosenthal and Strange (2008) and Combes *et al.* (2010) use characteristics of the subsoil since these factors influenced agricultural yields which represented a major reason for the establishment of early settlements. The difference with our paper is that size and density proxy for various observably equivalent agglomeration economies. However, the skill concentration exposes a specific externality. Robustness checks show that our instruments are neither driven by the historic size of regions, nor does additionally controlling (and instrumenting) for population size alter our findings.

A related approach in Acemoglu *et al.* (2001), is based on variation in settler mortality rates in the 16^{th} century as an instrument for the differing quality of institutions in order to explain countries' diverging development paths. Other papers that focus on the exploitation of natural resources, forced labor and institutions are those of Nunn (2008), Dell (2010) and Naritomi *et al.* (2012). In our approach industrial structures in the past provide a link to the current concentration of occupations and skills in regions within a single country. First nature advantages like proximity to the sea and railroads turn out to have a slight impact on the current wage distribution without, however, undermining the IVs' relevance.

We find a positive and highly significant effect of the concentration of face-to-face and analytical skills on the local wage level. This fact is documented by OLS as well as 2SLS estimations using a number of different specifications whereby we control for a multitude of

personal characteristics, sectoral affiliation, regional differences and permanent first nature advantages. Specifically, we find that an increase of one standard deviation in the local concentration of skills raises the average wage level in a region by about 10%. A large divergence between OLS and IV estimates indicates that endogeneity and measurement error in skills are substantial. Note, however, that none of our findings crucially hinges on the utilization of skill variables. We observe comparable wage effects from the concentration of managers, scientists and skilled technicians, i.e., those workers that use analytic and interpersonal skills most intensively. Since these skills are especially concentrated in large, urban regions, our paper also contributes to the explanation of the urban wage premium, cf. Glaeser and Maré (2001).

The paper most closely related to ours is Michaels *et al.* (2013). Those authors report for the US that in 1880 it was mainly manual/physical skills that were concentrated in urban areas, while nowadays, mostly interactive skills are located there. They provide a general equilibrium model with multiple regions, sectors, occupations and task/skills that explains this structural transformation of the economy. In essence, they conclude that the nature of agglomeration economies has changed over time. Due to falling transport and task trade costs, the benefits of concentrating physical production have diminished, while the concentration of interactive skills in densely populated regions has become more attractive. This argument is perfectly consistent with our findings. Yet not all densely populated regions exhibit a high concentration of interactive skills. Our paper adds to the picture that the transition of regions to the current equilibrium did not occur arbitrarily but was largely favored by the regions' sectoral supply and demand capacity in the past.[2] Moreover, we link the distribution of historical and current occupations to information about individuals, including their wages.

One of the few papers that also applies econometric methods to analyze the long-run economic development since 1872 in Brazil is Reis (2014). He also acknowledges the importance of both geographic factors and institutions (such as slavery) in explaining the persistent regional income inequality in Brazil. The paper documents the slow conversion of income per capita and labor productivity, i.e., the persistence of regional inequality patterns.

Finally, our approach is also related to papers on the dynamic aspect of agglomeration externalities and long-run economic development. Glaeser *et al.* (1992) examine city-industries over a period of 22 years to test whether industrial specialization, competition of diversity is more favorable for employment growth. The difference between our study and theirs is that they focus on the development of single localized industries and on employment growth directly, whereas we investigate long-term transitions from the industrial revolution up to the present day. Henderson *et al.* (1995: 1068) also acknowledge that dynamic externalities "lead to a buildup of local trade secrets". Again, the analyzed period in their study does not exceed 30 years.

[2] Further difference between our paper and Michaels *et al.* (2013) exist regarding the definitions of regions and skills. Michaels *et al.* (2013) use a different task/skill definition that serves to describe economic activity in 1880 as well as today. In turn, we define regions that are stable over time.

The remainder of the paper is organized as follows. Section 2 explains our estimation strategy. Section 3 reviews the historical background in Brazil and section 4 describes the utilized data. Section 5 contains the results and further robustness checks. Section 6 concludes the paper.

2 Estimating the agglomeration economies of skills

To identify the agglomeration economies of skills we start with two linear wage regressions. The two models incorporate skills and agglomeration economies in different ways. Both models include the log population size ($size_{ik}$) of region k where individual i works in order to account for the multitude of observably equivalent agglomeration advantages (Rosenthal and Strange 2004). Beyond that, we exploit skill measures that we derive from the occupation of each individual worker in our sample. The precise definition of our skill variable is contained in section 4.3.

In the first approach, we capture the effect of skills through the average value of the skill variable in region k. The estimation equation thus reads

$$w_{ik} = \beta_1 \overline{skill}_{ik} + \beta_2 size_{ik} + controls_i + \delta_{io} + \epsilon_{ik} \tag{1}$$

where w_{ik} is the log hourly wage of individual i and $controls_i$ represent workers' socio-economic characteristics provided in the census for 2010 (cf. section 4.1) namely: age and age^2, education, race, marital status, occupational position, gender, having a physical difficulty, being illiterate or a foreigner. The term $controls_i$ also includes some fixed effects to control for general wage differences between sectors and Federal States. Because the two variables of interest ($\overline{skill}_{ik}$ and $size_{ik}$) only have variation at the regional level, clustered standard errors are required (Moulton 1986). As in all of the following estimations, we use the sample weights provided by the Census.

Despite these individual controls, the average wage in a region may simply be higher because many white-collar workers are located there. This circumstance would go hand in hand with a high average skill score. Yet our aim is to distinguish the pure externality of the skill concentration from the average wage level which comes along mechanically with the occupational composition. Using occupation fixed effects δ_{io} (at the 2-digit level), we eliminate both the effect of individual skills as well as all other differences between professions that have an impact on wages, such as unobserved skills that are common within a peer group.

In the second model, we estimate how the hedonic price of skills varies with the local population size following Bacolod *et al.* (2009a). Thereby, our data and implied externalities can be directly compared with a related study from the US. In contrast to eq. (1), the skill value of individual i's occupation enters the equation directly, and only the coefficient of the interaction term γ_3 informs about whether the hedonic price of the *skill* differs among regions.

$$w_{ik} = \gamma_1 skill_{io} + \gamma_2 size_{ik} + \gamma_3 skill_{io} * size_{ik} + controls_i + \varepsilon ik \tag{2}$$

The term $controls_i$ represent the same variables as in eq. (1) above. However, we cannot include occupation fixed effects because the variable $skill_{io}$ only has variation at the occupation level. Since $size_{ik}$ is another aggregated variable, standard errors are clustered at the region-occupation level here.

Obviously, the interpretations of the coefficients of interest (β_1 and γ_3) are different in the two specifications. For ease of interpretation all independent variables (except for dummy variables) are centered on their weighted sample mean. Thus, in the second model, the sum $\gamma_1 + \gamma_3$ indicates the marginal wage effect that occurs if an individual i who lives in a region of medium size is employed in an occupation that demands slightly more of the specific *skill* category than the average worker in the economy is performing. The interaction effect γ_3 also shows if the hedonic price of the skill varies according to the size of the region. A positive coefficient indicates that performing a specific type of skill becomes more productive in the presence of *any kind* of other workers. In contrast, $\beta_1 > 0$ in the first model indicates that every worker, independent of his/her occupation, benefits from a higher concentration of a specific skill. In fact, eq. (1) measures a specific agglomeration advantage while eq. (2) expresses the general advantage of size for a specific activity. For this reason and because it also the path of least econometric resistance we prefer to work with the concentration of skills at the regional level as our main variable of interest. The specification with the interaction term between two endogenous variables is especially inappropriate when it comes to instrumenting for these variables, as we discuss below. Therefore, eq. (1) is continually enhanced during the rest the analysis.

Several threats to the identification of agglomeration economies arise in the OLS regressions. These can be summarized under the three well-known keywords: measurement error, omitted variables and reverse causality. First, large cities with the highest wages attract the best workers, just like Frank Sinatra said about New York: "if I can make it there, I'll make it anywhere".[3] An immediate consequence is that an influx of high-wage workers increases the average wage in the region, so that ever more high-wage workers are attracted, and so on and and so forth. Another problem with this selection of the best individuals to large regions is that, most likely, not all relevant wage determinants can be captured through education and the remainder observable personal characteristics. Thus wages may vary due to unobservables and as the study of Combes *et al.* (2008) or Ehrl (2014a) shows, this is especially true in populated regions. The problem in the present case is that workers with high wages and high unobserved skills are mainly also those who use analytical and interpersonal skills intensively. For these two reasons, the skill concentration is an endogenous variable and its OLS regression coefficient is generally biased.

In addition, measurement error may be present in each individual skill value. The values in the data are averages at the level of occupations extracted from representative surveys. However, it may be that the intensity of skills differs among regions. As argued above, workers who possess the best observable characteristics may use face-to-face skills more

[3] Eeckhout *et al.* (2014a) use the same quote of Frank Sinatra to motivate a matching model with complementarities between the highest and lowest skilled workers that results in spatial sorting of these types of workers to large cities.

intensively than people who formally have the same profession, but are employed in rural areas. Finally, the fact that a representative survey on skill data is absent in Brazil and we resort to data from the United States, may induce deviations from the true skill values, too. It is well known that due to this attenuation bias the estimated coefficients are closer to zero than their true value. The good news is that all three problems are mitigated at once by the use of instrumental variables.

From historical data from the years 1872 and 1920, we generate four instrumental variables so that it is possible to assess the exogeneity of the instruments with overidentification tests.[4] Building on eq. (1), we estimate the following three versions thereof in order to deepen our understanding of the underlying agglomeration mechanisms

$$w_{ik} = \beta_1 \overline{skill}_{ik} + controls_i + \delta_{io} \tag{3}$$

$$w_{ik} = \beta_1 \overline{skill}_{ik} + controls_i + \delta_{io} \begin{cases} \text{for highly populated regions in 1872/1920} \\ \text{for rural regions in 1872/1920} \end{cases} \tag{4}$$

$$w_{ik} = \beta_1 \overline{skill}_{ik} + \beta_2 size_{ik} + controls_i + \delta_{io} \tag{5}$$

where $\overline{skill}_{ik}$ and $size_{ik}$ are instrumented by one or more IVs. Running the regressions at the individual level is more suitable than at the regional level due to the multitude of worker-specific control variables. As an example, controlling for (mean) education at the regional level is inopportune, since the overall education level may be correlated with the size of the region, which makes education another endogenous variable. Instrumenting for one more endogenous variable is technically possible but not desirable because it complicates the interpretation of the results. This is why we avoid estimations with an interaction term of two endogenous variables, such as eq. (2). For the same reason, we prefer eq. (3) over eq. (5) where the term $size_{ik}$ partials out other types of agglomeration economies. The latter variable is indisputably as endogenous as the skill concentration measure. Nevertheless, as a robustness check, we control and additionally instrument for the current population size of regions.

Finally, we split the sample according to the historical size of regions in eq. (4). On the one hand, one can infer whether the agglomeration effects are heterogeneous, in a similar way to an interaction between the skill concentration and the size of the region. At the same time, one avoids the above-mentioned problems related to multiple IVs. On the other hand, we will argue below that the relation between the the historical IVs and the endogenous skill concentrations is significantly stronger if the region had a large population in the past. For this reason, we distinguish regions according to their size in 1872 and 1920, respectively, depending on the year of the instruments in use.

Estimations of eq. (3) to eq. (4) are primarily made with 2SLS. If more than one IV is used, estimation with GMM is more efficient (Baum *et al.* 2007). Except for eq. (2), standard errors are clustered at the regional level to account for spatial correlations across

[4] Given that the available number of available instruments is larger than one, it would also be technically possible to use more than one skill variable in the estimations. However, we do pursue this strategy due to multicollinearity concerns between the skill measures.

regions. Other cluster levels and combinations with the occupation-level are certainly possible but as in Rosenthal and Strange (2008: 381) we observe that in comparison to other clustering strategies, the clustering at the most aggregated, i.e., the regional level, is "a conservative approach in the sense that it has the greatest downward impact on the model test statistics". Clustered error terms complicate the calculation of the usual IV test statistics somewhat because most of the tests are based on i.i.d. errors in their original formulation. In the GMM estimations, the exogeneity of the instruments is tested by means of Hansen's J statistic which allows observations to be correlated within groups. A robustified version of Kleibergen's K statistic is used according to Finlay *et al.* (2013). This K statistic indicates whether the estimated coefficient is significantly different from zero assuming that the IV is exogenous. We also apply the widely recognized test for weak instruments by Stock and Yogo (2005) in its cluster-robust version provided by Olea and Pflüger (2013) which, however, is only applicable in the case of one endogenous variable.

3 Historical background

In 1872, Brazil was an undemocratic, rural, sparsely populated, slave-owning society and above all very poor. Its per capita income was only 1.8 times the subsistence level, similar to Malawi or poorer than Rwanda or Somalia today; and 16% of the population were slaves (The Maddison-Project 2010). When compared to other countries, its per capita income was half that observed in Argentina and just over a fifth of England's. Brazilians' life expectancy at birth was only 27.4 years and population density was 1.2 inhabitant per square kilometer (lower than contemporary Mongolia) (Mello 1984). In 1872, the great wave of non-Iberian immigration had not yet arrived with full force. Excluding those born in Portugal and Africa, there were only 120,000 foreigners in the country. The big wave of immigrants – many of whom possessed better skills than the average Brazilian – gained pace from the last quarter of the 19^{th} century onward. By 1920, Brazilian population had reached 30 million, of which 1.6m were foreigners or naturalized (Levy 1974: 79).

In 1872, the majority the almost 10 million inhabitants lived in municipalities close to the sea. The enormous size of the country and the *Serra do Mar* – a mountain range parallel to the Atlantic coast – have historically imposed high transportation costs on Brazil. Transport costs started to fall, however, in the last decades of the 19^{th} century with the expansion of the railway network, especially in the coffee-producing province of São Paulo. In 1870, there were only 678 km of railways; another 2504 km were completed over the next decade (Monasterio and Reis 2008). Up until 1930 the extent of the railroad network still had grown about tenfold. Reis (2014) shows that the current extension of railroads is greater but they still more extensive but still runs along the main lines from the past. Even so, the railroad system has never been of great importance for goods transport in Brazil. Marcondes (2012) shows that around the end of the 19^{th} century the largest part of interstate trade was handled via coastal shipping. Nowadays, truck traffic is more important than maritime commerce.

According to the 1872 Census, only 1.5 million people were literate. Stolz *et al.* (2013) also show that numeracy of Brazilians was quite low in the 19^{th} century. This dismal social situation persisted for a long time. According to Chaudhary *et al.* (2012), only 12% of school-age children were enrolled in primary school in 1910, while at the same time 80% of children in the UK, Germany and the US were studying. Consequently, the number of teachers – 7.2 per thousand inhabitants – was also extremely low in Brazil , compared to 58 in the US, 33 in Argentina and 15 in Chile (Kang 2010: 43). In view of these circumstances, the wealthy have traditionally relied on private schools and private teachers. With the advent of industrialization, the local elites supported the expansion of mass schooling, though only in regions where there was a demand for skilled workers (Chaudhary *et al.* 2012). It was not until the end of 20^{th} century that primary education was universalized. Likewise, the situation in higher education was also precarious. Some Law, Medical and Engineering schools were established throughout the 19^{th} century but in 1872 there was still no university in Brazil.[5]

First steps of Brazilian industrialization – mostly light industries with imported machinery – took place in the last decades of 19^{th} century. Manufacturing activities in 1872 were labor intensive and small-scale; it is from the 1920s on that full blown industrialization took place. The primary sector's share in the GDP fell from 38% to 9% between 1920 and 1980. In the same period, the manufacturing industry grew from 12% to more than a third of the GDP (Reis *et al.* 2002: 248). Economic growth rates of the Brazilian economy have varied widely since 1872. It is estimated that the income per capita was stagnant in the last two decades of the 19^{th} century. Between 1900 and 1930, income grew at a rate of 1.5% per year and accelerated to an impressive 3.3% a year between 1930 and 1980. Again, the following two decades were called lost decades (average growth 0.2 %) and, finally, moderate growth has returned in the last 15 years (Reis *et al.* 2002).

4 Data

Our primary data sources are the official Censuses from 1872, 1920 and 2010. Some aggregate variables are generated from data administrated by the National Brazilian Institute for Geography and Statistics (IBGE) and from the Institute of Applied Economic Research (IPEA). Finally, we use a mapping between current Brazilian and US occupations in order to enrich our data set with information about workers' skills from the O*NET.

4.1 Censuses from 1872, 1920 and 2010

The first Brazilian Census undertaken with reliable methods and a complete coverage of the Brazilian territory was carried out in 1872 (Botelho 2005). For the purpose of the present study, we extracted the citizens' parish of residence and their professions. Due to

[5] Public universities were created in Rio de Janeiro and Curitiba in the first two decades of the 20^{th} century. Until then, members of the local elites used to study at European universities (Fávero 2006).

the political instability caused by the end of the empire, the quality and comprehensiveness of the following Census in 1890 is poor (Monasterio and Reis 2008).

The next Census from 1920 provides reliable data. Again, we make use of the residential and occupational information of the population. The single categories are not directly comparable between 1872 and 1920 but an exact match between all categories is not even required by our identification strategy.[6] Besides, only two key groups of historical trades are of interest here: liberal and manufacturing professions. The former are divided into more categories than in 1872, while the manufacturing sector has undergone severe structural changes. To minimize consistency problems all of the liberal professions as well as the professions in the manufacturing sector are combined into two different 'super-categories'. In any case, we regard the very fact that industrial production takes place as more important than the question of which type of manufactured product is actually produced by whatever occupation.

The most recent Census stems from the year 2010. In contrast to its earlier versions, it is separated into two parts. Besides the obligatory, basic demographic survey, there is a more detailed questionnaire that a random sample of households has to respond. In 2010, it covered 10.7% of all households, or an equivalent of 6.2 million homes with 20.6m individuals. The selection of households corresponds to a stratified sample of five size types of municipalities, see IBGE (2010) for further details. The share of surveyed households is inversely proportional to the size of the municipality.[7] Therefore, we apply the sample weights provided by the IBGE to our calculations and obtain representative results for the entire population and the actual sectoral composition of the country.

According to our identification strategy, the Censuses from 1872 and 1920 are used to generate historical instruments, while the Census from 2010 provides the following information about workers and their wages. We focus on individuals between the age of 15 and 65 who declare themselves as having a job, work a positive number of hours and have a labor income larger than zero. We can thus calculate and use the hourly wage as the dependent variable. In addition, we exclude civil servants and members of the armed forces since their wages do not vary by region and thus are unable to reflect agglomeration economies. The information contained in the sample is very extensive. For our purposes, only those socioeconomic characteristics that affect the wage are of interest; such as: the place of residence, the sectoral affiliation of the individual, gender, age, nationality (foreign, native), race (5 groups), marital status (4 groups), having a physical difficulty and occupational position (employer, formally employed or not). The occupation, the years of schooling, and as previously mentioned the wages and hours worked are of particular importance. According to the different stages of formal education in Brazil, we construct dummy variables for the following five groups: (1) less than 4 years of schooling (incomplete primary education); (2) 4 to 7 years (primary and lower secondary education, *Ensino Fundamental I*); (3) 8 to

[6] In 1920, these trades were classified into 48 different categories altogether, whereas in 1872 36 categories were distinguished.

[7] For example, in municipalities with up to 2,500 inhabitants, 50% of all households covered by the more detailed survey. In the largest cities with over 500,000 inhabitants, the proportion is merely 5%.

10 years (higher secondary education, *Ensino Fundamental II*); (4) 11 to 14 years (high school graduated, *Ensino Medio*); (5) 15 or more years (college or university graduated).

Finally, we complement these Census data sets with maps from the IBGE. Using GIS software, we calculate the shortest distance from the centroid of an area to the coastline. The IPEA provides data on the area of the municipalities (in 2010) and on the number of train stations. The latest information available on the latter is from 1995. The large share of the informal sector in Brazil complicates a clear distinction between working or employable population even today. The following calculations of the size of regions in 1872, 1920 and 2010 are thus based on the total population.

4.2 Generation of Minimum Comparable Areas (AMCs)

The changing demarcation of municipalities poses the greatest difficulty in the inter-temporal applicability of the Census data from 1872 to 2010. Municipalities are the smallest units of the Federation endowed with administrative autonomy. Their number increased from 624 in 1872 to 5,570 today.[8]

For our analysis, however, we need a stable, i.e., a comparable spatial delineation of regions over time, the so-called 'Minimum Comparable Areas' (AMCs). Such a task has previously been undertaken in Reis *et al.* (2011). However the period used in their study ends in the year 2000, when 'only' 5507 municipalities existed. Further methodological differences and the details of the present approach are described in Ehrl (2015). The basic idea is to combine current municipalities so that the aggregates are exactly consistent with the ancient borders of municipalities, taking all combinations and divisions of municipalities over the last 130 years into account.[9] The resulting number of 479 different AMCs is higher than the one in Reis *et al.* (2011), suggesting that the retracing of municipalities' family trees has become more accurate. This is good news, because the identification of agglomeration effects is based on variation among AMCs.

Figure 1 provides an overview of the result. It is notable that the further away from the sea, the larger the area of the municipalities. This pattern reflects the progressive populating process of an ex-colony emanating from the coast. The vast territory of many AMCs is thus not primarily related to the aggregation procedure but to the dimension of the municipalities in the interior. Four of them even have an area of over 100,000 sq km, which is larger than the territory of Portugal, for example. For areas of such dimension, it obviously makes no sense to estimate agglomeration economies, because they arise from the interaction between individuals within localized labor markets. It is thus necessary to limit the study to those AMCs that have the size of a coherent labor market. A reasonable number is 2,500 sq km. This corresponds to the area of a square with sides of 50 km, i.e.,

[8] Reasons for the enormous increase in the number of administrative units are the expansion of the population and the economy and the political decentralization during the 1960s and 1970s (Reis *et al.* 2011).

[9] In a few cases the municipalities are aggregated across two Federal States. Some Federal State did not yet exist in 1872, so that an aggregation of these states is also required. In total, we end up with 16 different states.

an acceptable maximum distance to commute. Besides, this is also the average area of municipalities in the US.

Because of these arbitrary restrictions and in view of the remaining AMCs (cf. the dark areas in figure 1) further comments are warranted. First of all, estimating the preferred specification (eq. (3)) in the sample with all AMCs – despite the concerns described above – yields qualitatively the same results that we present in the main text. Also, further tightening of the restriction on the AMCs' areas does not change the fundamental insights of the paper. Considerations regarding the instruments also point out that the the area size restriction is preferable. Some AMCs are only vast because the municipality had a large territory in 1872. Even though nowadays a large city like Goiania, Brasília or Porto Alegre is contained in this vast AMC, the data does not reveal if and how far other settlements around that city influenced the value of the historical instruments. Obviously, Brasília did not even exist in 1872 or 1920 and thus the concentration of historical trades in whatever settlements existed within this AMC should not be used to predict the concentration of skills in Brasília in 2010. In any perspective, the accuracy of the instrument deteriorates undesirably with the AMC's area. The descriptive figures in section 5.1 show that despite the area restriction, there is still extensive variation in population size and in the remainder of the essential variables exploited in this paper. These evaluations make us confident that the restriction to AMCs with an area of less than 2,500 sq km ultimately strengthens the credibility of the estimated agglomeration externalities.

Figure 1: Delineation of AMCs and municipalities

Notes: The figure shows the territory of Brazil to date. The fine lines indicate the borders of the municipalities in the year 2010, whereas the bold lines mark the frontiers of the aggregated Minimum Comparable Areas (AMCs) for the period 1872–2010. Note that the territory of the Federal State Acre is excluded from our analysis because it was part of Bolivia in 1872. AMCs marked in dark are those with an area of less than 2,500 sq km. The observations from these AMCs define our sample.
Source of the GIS delineation: IBGE.

4.3 Skills

Skill measures are the result of a division and systematization of individual activities in everyday work life. On the basis of the frequency and the intensity with which a specific skill/task is performed, occupations, and eventually workers, may be compared along a small number of skill dimensions. A further advantage of skills is that the distinction of workers according to their actual activities is more meaningful for some issues of economic analysis than, for example, their formal education. Skills/tasks recently found their way into labor economics and other related research fields through the work of Autor *et al.* (2003). They argue that technological progress in the form of IT, computers and automation can replace routine and manual activities, whereas it complements analytical and cognitive skills. Consequently, one can also expect workers' skills to play an important role in the analysis of agglomeration economies.

In Brazil, no workforce survey exists that assesses the skill/task requirements of workers in their occupations. Thus the only way to make progress is to adopt data from another country.[10] One advantage thereof is that we can use established definitions which make this study comparable to existing work. We tried several skill definitions, but throughout the study, we confine ourselves to two types of skills, to avoid duplicities and thus to save space. Following Acemoglu and Autor (2011) and Firpo *et al.* (2011), we distinguish between analytical and "face-to-face", i.e., interpersonal skills.[11] To make the skill scores comparable among one another, we standardize them to a mean of 10 and a standard deviation of 1 in the occupation data. Section A in the appendix provides a justification for the choice of these two skill measures alongside some numbers and graphs that relate the distribution of skills, occupations and wages. The observed pattern resemble those in the US or Germany.

5 Results

5.1 Descriptives

5.1.1 Concentration of skills

The current section provides an overview of the spatial distribution of the skill variables. This is interesting, not only because of the new definition of AMC regions, but also to verify whether Brazil exhibits patterns comparable to those of other countries.

The vertical axis of the graphs in figure 2 denotes the concentration of the skill measures. This concentration is defined as the population-weighted average of a skill measure in the AMC. These averages vary between 9.2 and 9.9 in the case of analytic skills. It already

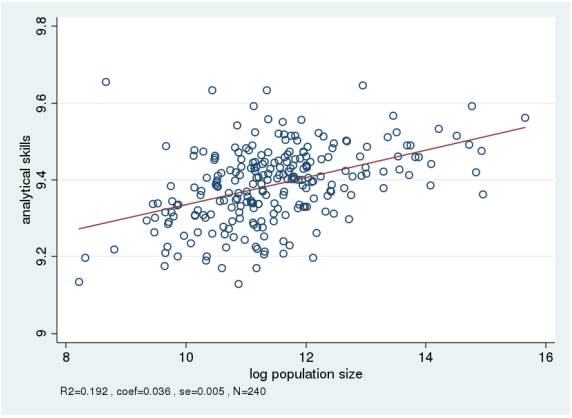
[10] The assignment of an occupation characteristics survey to another country has been made previously. Andersson *et al.* (2014), for example, transfer a German task classification into Swedish data, and Ehrl (2014b) relies on the same mapping between Brazilian jobs and US skills in a study about offshoring.

[11] Bacolod *et al.* (2009a) rely on a less recent survey (the DOT) for their skill definitions but use a similar category which they call "people skills".

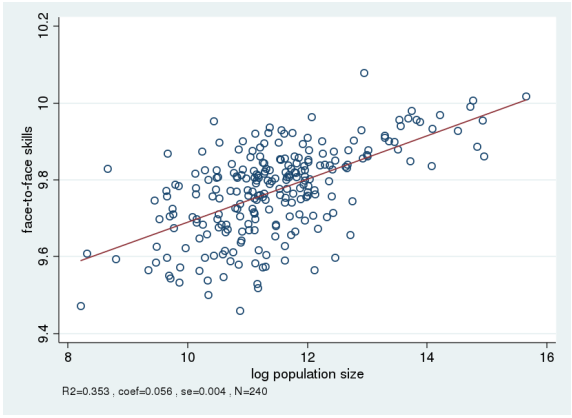
becomes evident that there are considerable differences in the vocational orientation of regions. Each scatter plot additionally reports the outcome of a simple unweighted OLS regression of the skill concentration on the log population per AMC. Both graphs show a clear pattern. The larger the size of the region in terms of population, the more analytical and face-to-face skill are performed in the local economy.[12] The strength and explicative power of the regions' size is somewhat higher for the concentration of face-to-face skills. Nevertheless, the robust standard errors confirm that both positive correlations are highly significant.

Figure 2: Spatial concentration of skills – AMC means

(a) analytical



(b) face-to-face



Notes: The circles in each graph represent the AMCs' skill averages and their log population size. Only AMCs with an area smaller than 2,500 sq km are part of the sample. The results from the corresponding (unweighted) linear regression are reported below each graph.

To get an idea of the dimension of the skill differences, consider the following stylized examples. São Luís do Quitunde in the Federal State of Alagoas in the North of Brazil has an average analytical skill value of 9.2 which is equivalent to a worker in a paper factory. In Florianópolis, in contrast, the capital of the State Santa Catarina in the South which has one of the highest analytic skill concentrations, an average occupation has the equivalent of a mechatronics technician or human resource analyst (9.6). In Areias (São Paulo), one of the smallest AMCs with only 3696 inhabitants, about 40% of the population make a living from agriculture and another 33% work in the construction or retail trade. The face-to-face skill index in Areias has a value of 9.47 which roughly corresponds to a worker in the construction sector. On the other side of the face-to-face skill distribution cities like Recife, Belo Horizonte and Rio de Janeiro have a value of about 10. Artists, like musicians or art directors, have just such an interactive skill score.

A similar overview for Brazil is provided in Maciente (2013). Although the definitions of the skills are different, namely less aggregated, and their alignment is by size groups of municipalities, it becomes clear that conclusions remain the same, that is to say, skills related to discipline, independence, attention or communication are more concentrated in

[12] Analog figures with the spatial distribution of manual and cognitive skills are contained in appendix A.

large municipalities.

5.1.2 Remuneration of skills

As a further motivation of our main analysis, we consider how the concentration of skills is related to the average wage level of a region. This time, both the plots and the univariate OLS regressions in figure 3 are weighted by the value of the population in the region. This representation is more important because it will be re-encountered in various wage regressions hereafter, which are estimated at the level of individuals.

As conjectured, regions where a disproportionately large amount of analytical and interactive skills are executed have a substantially higher wage level, cf. figure 3. On the other hand, a concentration of manual skills, which can be particularly found in rural areas, is related to a lower overall wage level, cf. figure A.3 in Appendix B. Some AMCs are quite distant from the estimated regression line. These AMCs have an average of above 9.8 for interpersonal skills and at least 9.5 for analytical skills but at the same time they exhibit a very low wage level. A separate consideration reveals that these regions are predominantly rural and the high score in analytical skills stems from a large number of agricultural producers, agronomist, food scientists, etc. A robustness check in section 5.6.2 addresses this peculiarity.

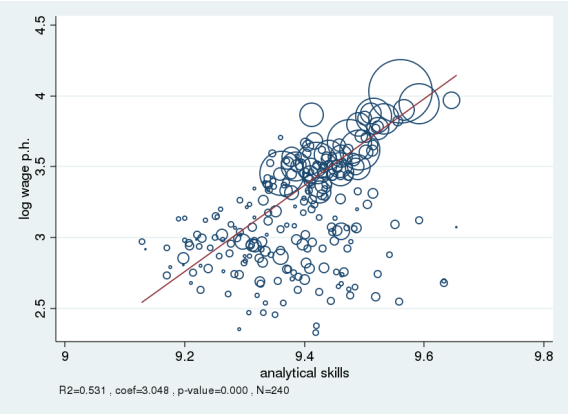
How can those increasing regression lines be interpreted? On the one hand, one can imagine that the skill content of each occupation changes, without altering the region's occupational composition per se. This is a counterfactual in which the daily work routine of every person is changed but admittedly it is less clear what the increase of skills by a certain value exactly means for each different profession. On the other hand, one could consider how the composition of occupations would have to change, to raise the value of the skill concentration in an AMC by a certain percentage, assuming the skills in each profession remained the same.

The difference of +1.0 in interpersonal skills equals the difference between a mechanical engineer or a wood technician and a manager (11.15 on average). So if everybody in a certain AMC would use this much more face-to-face skills, each wage would rise by 258%. This is certainly a high and unrealistic number, to the same extend as the large increase in interpersonal skills is unrealistic. Scaling the hypothetical increase in the average skill measure down to a less striking number of +0.01 thus implies average wage gains of 2.6%. Put differently, to raise the average face-to-face skill value of an AMC with a population equal to the mean (790,000) by 0.01, the region would have to experience an inflow of about 6,300 managers which corresponds to an increase of 8 managers per 1000 inhabitants.

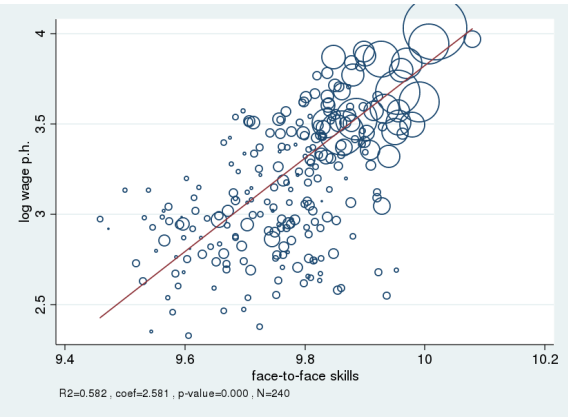
The positive correlation in figure 3 is not yet overwhelming evidence for the existence of agglomeration economies. After all, professions that require a high level of analytical and interactive skills are well paid and so it is rather mechanical that the average wage level of a region is correlated to the concentration of skills. Before we extend the interpretations, we want to consolidate these current findings.

Figure 3: Correlation between wages and concentration of skills – AMC means

(a) analytical



(b) face-to-face



Notes: The circles in each graph represent the AMCs' analytical and interpersonal skill and log wage averages using population weights. Only AMCs with an area smaller than 2,500 sq km are part of the sample. The results from the corresponding (unweighted) linear regression are reported below each graph.

5.2 OLS estimations

To make the descriptive observations about agglomeration economies in the previous subsection more reliable, we propose two extensions. For one, we want to control for other wage determinants, such as education, age etc. It is thus more appropriate to consider wage regressions at the individual level. For another, we distinguish the effect of a spatial concentration of similar skills from other types of agglomeration economies, which, generically, may be captured by the size or density of the region (Rosenthal and Strange 2004).

The first column in table 1 shows the results from the estimation of eq. (1) where β_1 is set to 0, i.e., the effect of population size alone. The next two columns show the effect of the regional concentration of interactive skills without controlling for general agglomeration economies (β_2=0). Columns (4) and (5) include both of these variables and column (6) presents the results from the estimation of eq. (2). All of these estimations include our standard worker-specific controls as well as sector and state effects. Columns (3) and (5) additionally include occupation dummies. Virtually all of these control variables are highly significant and their coefficients are as expected. We report coefficients only for the most common and important variables. Wages increase monotonically in the education level, there is a slightly bell-shaped relation to the workers' age and males earn more than observably equivalent females.

The elasticity of population size, when other worker-specific characteristics are controlled for in column (1) is equal to 7%. This elasticity is slightly larger than in Bacolod *et al.* (2009a) who report a coefficient of 6.6%, even though we control for more and more detailed personal characteristics and our $R^2 = 0.34$ exceeds their R^2 of 0.22. The elasticity with respect to size falls when the concentration of face-to-face skills is added to the estimation. In line with common intuition this result suggests that the inclusion of a specific

Table 1: Wage regressions on face-to-face skill measures

	Dependent variable: log wage p.h.					
	(1)	(2)	(3)	(4)	(5)	(6)
skill:		AMC-level	AMC-level	AMC-level	AMC-level	individual
skill		0.748***	0.661***	0.177	0.119	0.166***
		(0.073)	(0.071)	(0.143)	(0.135)	(0.009)
log(size)	0.070***			0.061***	0.057***	0.068***
	(0.004)			(0.010)	(0.010)	(0.004)
skill*log(size)						0.043***
						(0.004)
2.educ	0.011*	0.010*	0.009	0.010*	0.009*	0.013**
	(0.006)	(0.006)	(0.005)	(0.006)	(0.005)	(0.005)
3.educ	0.162***	0.163***	0.140***	0.161***	0.138***	0.159***
	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)
4.educ	0.353***	0.353***	0.274***	0.352***	0.273***	0.335***
	(0.010)	(0.011)	(0.008)	(0.011)	(0.008)	(0.009)
5.educ	1.140***	1.143***	0.819***	1.137***	0.816***	1.063***
	(0.027)	(0.029)	(0.019)	(0.029)	(0.018)	(0.023)
age	0.012***	0.012***	0.011***	0.012***	0.011***	0.012***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
age^2	−0.000***	−0.000***	−0.000***	−0.000***	−0.000***	−0.000***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
male	0.224***	0.222***	0.185***	0.224***	0.185***	0.235***
	(0.009)	(0.009)	(0.008)	(0.009)	(0.008)	(0.009)
occ. dummies	✗	✗	✓	✗	✓	✗
Constant	2.483***	2.432***	3.179***	2.477***	3.213***	2.516***
	(0.038)	(0.047)	(0.065)	(0.038)	(0.059)	(0.039)
Obs.	1,293,046	1,293,046	1,293,046	1,293,046	1,293,046	1,293,046
R^2	0.342	0.339	0.372	0.342	0.374	0.348

Notes: The skill variable in columns (2) to (5) is calculated as the weighted AMC average of face-to-face skills, whereas in column (6) the individual face-to-face skill value is used in the regression. Therefore, standard errors in brackets are clustered at the AMC-level in columns (2) to (5) and clustered at the AMC-occupation-level in column (6). Besides the variables shown in each row, the regressions also include dummies for sector, Federal State, occupational position, race, marital status, and whether or not the person is a foreigner, illiterate or has a physical deficiency. Regressions are weighted by the official sample weights provided by the Census. * denotes significance at ten, ** at five and *** at one percent level.

agglomeration effect diminishes the strength of the general effect of size. However, the change here is small.

The coefficients of the skill concentrations in columns (2) and (3) are again positive and significant, albeit much smaller in magnitude than in the respective univariate regressions at the AMC level reported in figure 3. The coefficient in column (3) suggests that if the face-to-face skill level of a region is increased by 0.1, equal to the variable's standard deviation, wages increase by 6.6%. This coefficient increases by a magnitude of 13% when occupation-specific effects on the wage are not eliminated. Including both population size and skill concentration shows that only the former shows a positive and significant coeficient. These simple OLS regressions suggest that the generic agglomeration economies induced by the region's population size dominate the effect of a concentration of interpersonal skills.

The last column in table 1 presents the estimation of eq. (2). The results also suggest that there are sizable agglomeration economies. The elasticity of wages with respect to population – evaluated for workers with face-to-face skill scores equal to the sample mean – (e.g. an actuary with a score equal to 9.8) is equal to 6.8%. The positive interaction coefficient in the third row indicates that the generic agglomeration effects rise, the higher the interpersonal skills of the worker and the larger the area's population are. A worker with an interpersonal skill value of a standard deviation above the mean (executive directors for example who have a score of 10.9) has a by 4.1% higher wage elasticity. Altogether, these workers experience wage gains of 11.1% when the region's population doubles, which corresponds to an increase of 63% compared to the elasticity of book keepers. This result suggests that interpersonal skills are necessary in order to capitalize the diverse agglomeration economies in large regions. These observations are in line with the literature on the importance of interaction in urban environments cf. Fujita and Thisse (2002). In particular, our results are also quantitatively similar to those in Bacolod *et al.* (2009a: 145). Using US data, those authors obtain a (plain) wage elasticity with respect to population of 4.4% and workers "with the ability to interact" also have a more than 50% higher elasticity.[13] The extent to which the entire population or only some groups benefit from the spillovers is investigated in subsection 5.5.

5.3 Basic IV results

Our instrumental variable strategy intends to identify the unbiased and causal effect of an agglomeration of interpersonal skills on wages. The assumption underlying the identification strategy is that the occupational structure of a region 140 or 90 years ago has been shaping the further economic development of that region until the present day. In particular, a high concentration of industrialists and liberal professions has fostered a positive long-run development which has resulted in a conglomeration of high-skilled jobs, which largely require those interactive (and analytical) skills.

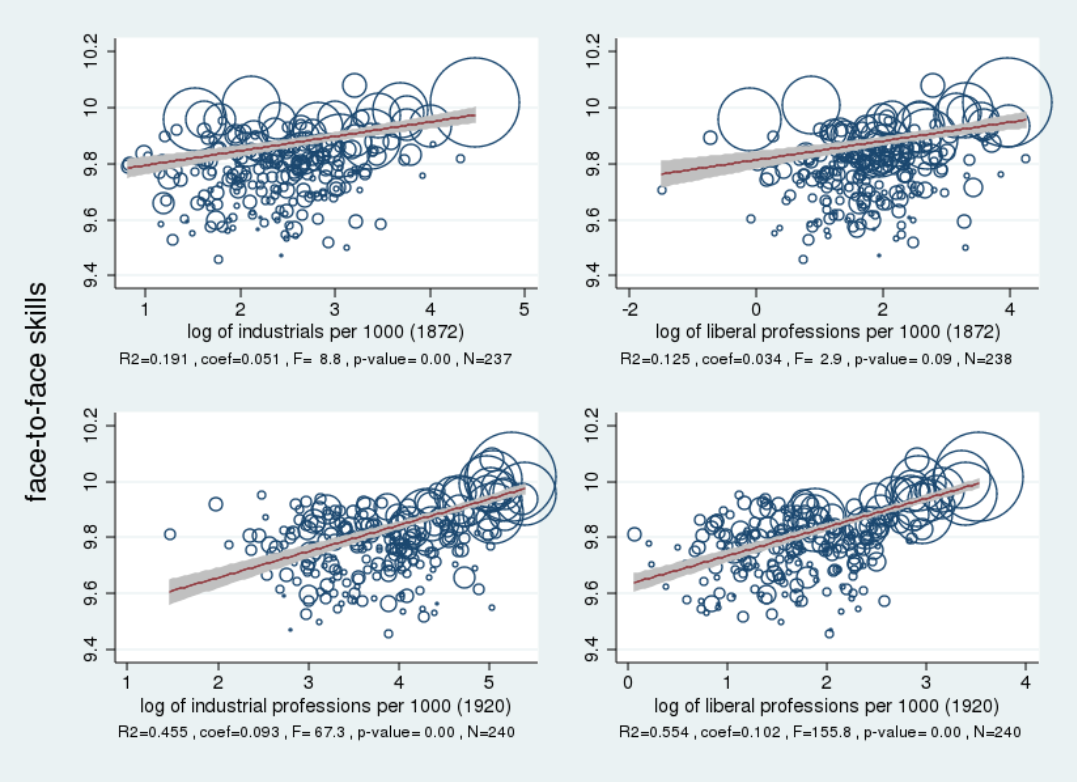
Section 3 described how both education and manufacturing production in factories were scarce and hence quite valuable at the turn of the 19^{th} century. Moreover, it is likely that lawyers and judges, who are also part of the liberal professions served to stabilize the fragile institutions, maintain law and order and guarantee property rights, all essential conditions for investment and economic growth (Acemoglu *et al.* 2001). The composition of the local economy was much more important than nowadays because railroads had still hardly been built and hence overland transport costs were high. This constellation favored economic development in those regions where the seed for knowledge and industrial production was already sown, whereas regions shaped by agriculture and slavery were likely to remain locked-in and without impulses towards modernization.

Figure 4 illustrates our identifying assumption. The upper left graph of figure 4 shows

[13] Using log population density instead of population size produces qualitatively very similar results. The coefficient of log population density is around 0.06 in a regression without skill variables comparable to column (1) in table 1. Due to space constraints and in order to provide comparable results to the prior study by Bacolod *et al.* (2009a), we only present the findings using log population size.

that the proportion of industrialists in 1872 explains 19% of the spatial dispersion of interpersonal skills in 2010. The coefficient of this linear regression – stated below the graph – is highly significant and indicates that doubling the number of industrialists in a region is associated with a 0.05 (or $\frac{1}{2}$ standard deviation) higher face-to-face skill index. The relation is much weaker regarding the share of the liberal professions in 1872. In this linear regression, the instrument has to be classified as weak. This assessment definitely changed 48 years thereafter. Now the explained percentage of 55% is also higher than that of the manufacturing sector. Regarding analytical skills a very similar picture emerges; compare supplementary figure B.1.

Figure 4: Correlation between face-to-face skill mean and historical trades



Notes: The circles in each graph represent the AMCs' interpersonal skill average and each one of the four instrumental variables in the sample. The results from a weighted linear regression of each of the IVs on the face-to-face skill concentration are indicated below each graph.

It is certainly impossible to pin down a single channel through which history has shaped the present. Yet figure 4 suggests that our instruments go a long way towards explaining the long-run industrial development of regions. The data also show that the less far we look into the past, the stronger becomes the relation to the previous occupational composition, albeit the structure of the economy has still changed significantly over the last 90 years. The following tables deepen the interpretation and the exploration of the relevance and exogeneity of our instrumental variables.

Table 2 show the results from the estimation of eq. (3) where the concentration of face-to-face skills is instrumented.[14] The set of exogenous individual and sectoral control variables

[14] Results for the concentration of analytic skills are very similar to those for interpersonal skills in the

is the same as in previous estimations, cf. table 1. Again, by means of occupation fixed effects we also account for the fact that the actual professions of workers also affect wages. The first column repeats the OLS coefficient for comparison. In the next four columns, each one of our historical IVs is used. Then, both IVs from 1872, and 1920, respectively are used jointly. The last two columns show the results when all four instruments are applied simultaneously. In the lower part of the tables, the corresponding coefficients of the IVs from the first-stage regressions are presented along with other statistics that allow for the assessment of the IVs' performance.

By and large, the same picture emerges as described by means of the four simple regressions in figure 4. Considered separately, all instruments' coefficients are significant at the 1% level in the first stage regressions. The magnitude of these coefficients is similar to the one identified in the univariate regressions beforehand, and the interpretations of the coefficients apply equally. When considered jointly, only one or two of the instruments remain significant in table 2. Based on the univariate scatter plots, we were already able to foresee which one dominates. For 1872, the concentration of industrialists has the strongest effect on the concentration of analytic skills, whereas for 1920, the concentration of liberal professions is the more powerful instrument.

Table 2: Basic IV regressions – face-to-face skill concentration

	Dependent variable: log wage p.h.								
	OLS	IV (1)	IV (2)	IV (3)	IV (4)	1872 IVs	1920 IVs	all IVs	all IVs
face-to-face	0.661***	0.765***	0.898***	1.066***	1.064***	0.854***	1.064***	1.086***	1.004***
skill conc.	(0.071)	(0.173)	(0.159)	(0.108)	(0.092)	(0.153)	(0.092)	(0.087)	(0.086)
	1.-stage statistics								
1872: log of		0.062***				0.045***		0.012	0.012
ind. per 1000		(0.012)				(0.013)		(0.012)	(0.012)
1872: log of			0.044***			0.017*		0.005	0.004
liberal per 1000			(0.010)			(0.010)		(0.009)	(0.009)
1920: log of				0.087***			0.022	0.019	0.020
ind. per 1000				(0.009)			(0.014)	(0.013)	(0.014)
1920: log of					0.093***		0.075***	0.069***	0.068***
liberal per 1000					(0.006)		(0.011)	(0.012)	(0.012)
1. R^2-part.		0.182	0.152	0.409	0.487	0.191	0.493	0.507	0.462
weak IV: F		26.900	20.750	106.100	213.100	18.860	105.800	53.390	44.400
weak IV: τ=5%		37.420	37.420	37.420	37.420	19.040	13.320	20.960	20.810
weak IV: τ=10%		23.110	23.110	23.110	23.110	12.310	8.947	12.730	12.640
weak IV: τ=20%		15.060	15.060	15.060	15.060	8.462	6.418	8.214	8.160
weak IV: τ=30%		12.040	12.040	12.040	12.040	6.994	5.453	6.558	6.518
K						1.316	12.730	7.432	5.207
K-p						0.251	0.000	0.006	0.023
Hanson J						1.194	0.001	3.598	3.474
Hanson J-p						0.275	0.972	0.308	0.324

Notes: The regressions control for a quadratic in worker's age, dummies for occupation, sector, Federal State, education level, occupational position, race, marital status, and whether or not the person is a foreigner, male, illiterate or has a physical deficiency, is as specified before. All regressions are weighted by the Census population weights. The estimation in the last column additionally includes the 1^{st} nature advantage proxies. Standard errors in brackets are clustered at the AMC-level. The number of observations varies slightly due to few missing values of the IV but it lies above 1,290k in all of the estimations. * denotes significance at ten, ** at five and *** at one percent level.

The main novelty from these IV regressions compared to the OLS regressions in table

main text and can be consulted in the appendix table B.2.

1 is that the effect of the instrumented skill variable is now larger. The effect of the interpersonal skill concentration lies between 0.8 and 1.1. We conclude that measurement error and endogeneity bring about some substantial bias of the agglomeration effects of skill concentrations estimated by OLS. Using instrumental variables, the skill concentrations have a significant and positive effect throughout all estimations here. The weak IV statistics based on Stock and Yogo (2005) demonstrate that these coefficients, particularly when two or all four IVs are used, are estimated with a bias of less than 5–10%. The K statistic provides another way of testing the relevance of the instruments. It also reflects the critical view of the 1872s instruments. For the 1920s IVs and the joint estimations, however, it clearly underscores their high predictive power. Hansen's J statistic, however, offers a uniform picture. Wherever it is possible to perform the overidentification test, the IVs' exogeneity clearly cannot be rejected.

The last column in table 2 repeats the estimates with all four IVs but additionally includes controls for 1^{st} nature advantages. Of course, the distance to the nearest coastline has not changed over time. But also the number of railway stations after the peak of its construction activity has remained relatively constant until the present day. If settlement decision by industrialists and people with liberal professions as well as the long-run economic development up to the skill distribution in today's industries were substantially affected by those 1^{st} nature advantages, then the coefficients and in particular the explanatory power of our IVs should differ clearly from the previous estimates. Note that these estimations represent our preferred specification because all IVs are used simultaneously and we control for the maximum number of exogenous variables.

As expected, the distance to the coast shows a negative but insignificant sign in the second stage regression and the number of railroad station has a positive and significant correlation with wages (not shown in the tables). Yet the results remain almost unchanged. Both the first stage coefficients of our IVs and the instrumented coefficient of the skill variable are close to the estimates in the previous column. Even the partial R^2 of the IVs in the first stage regression is not changed much. We therefore conclude that distance from the coastline and the railroad transport system may represent a productivity advantage even today, but their existence has not affected the distribution of historical professions in a crucial manner. Thus, the exogeneity of our IVs is confirmed once again.

The coefficient in our preferred estimation indicates that an increase in the face-to-face skill index by one standard deviation (equal to 0.1) results in an average wage increase of 10%. To induce this change of +0.1, a region with a population and skill index equal to the sample mean would require an inflow of +88 managers per 1000 inhabitants, for example. This true effect of interpersonal skills is thus much smaller than the one conjectured from the simple OLS regression without any control variables in figure 3.

5.4 IV estimation with different size groups

This section examines whether there exist heterogeneities of the identified agglomeration economies with respect to the regions. OLS regressions with interaction terms so far reveal

that the wage externalities increase with the size of the region. Since current population is an endogenous variable, too, pursuing this strategy further would mean to instrument two endogenous variables plus their interaction term. Estimating and interpreting such a model with interaction of instruments is beyond the scope of the paper. A feasible possibility is to split the sample into groups based on the size of the regions and to estimate the effects within those separate subsamples.

At the same time, and even more importantly, we can perform another investigation with this regional distinction. Our hypothesis is that the long-run development of regions crucially depended on their initial size. In particular, we rely on *New Economic Geography* theories, which address the attractiveness of agglomerations and the process of their formation. Up to here, our historical instruments show that the concentration of knowledge and manufacturing has led to a continuous modernization and finally to a higher concentration of high-skilled occupations. However, it could be that this supply-side agglomeration is just a necessary but not a sufficient condition for a sustainable growth path. In line with the initial quote by von Thünen, New Economic Geography teaches that circular causality between demand and supply side promotes the growth of regions and guarantees their continuous attractiveness for workers and firms. This means that if a region was blessed with a sufficiently large number of customers for fabricated products and people who make use of the locally available knowledge, this region would experience higher growth rates than small regions. Consequently, regions with such favorable characteristics in the past should have a greater concentration of high-skilled occupations that predominantly require analytical and interpersonal skills.

5.4.1 Definitions of AMC groups according to population size

Following the theoretical arguments above, we divide the AMCs according to their population size in 1872 or 1920 into two groups each. AMCs in the "Large" group have a population above the median in the distribution of AMCs. Consequently, the other half of the AMC in the "Small" group has a lower population. We prefer this specific division because dividing the sample according to the population weighted mean, for example, gives approximately the same number of people in both groups, but the number of AMCs in the "Large" group becomes very low. Given that the number of AMCs with an area of less than 2,500 sq km amounts to no more than 240 and the main variables of interest only have variations between AMCs, the division according to the median is the best option that allows for a reasonable identification of the effects in both groups.

As the two maps of Brazil in supplementary figure B.2 show, the spatial distribution of AMCs across the two size groups is quite random. There is only a slight tendency that more AMCs in the South of the country are part of the *Large* group. Since the following regressions include Federal State dummies and the distance to the coastline, possible imbalances in these respects are eliminated. Moreover, the distributions of four IVs exhibit only small differences across the two agglomeration groups as the kernel densities in figure B.3 and figure B.4 in the Appendix B show. Contrary to expectations, the top of

the density distributions in 1872 of the *Small* size group are slightly skewed to the right. This pattern is reversed in 1920. Although it is rather natural that a higher concentration of liberal professions are to be found in more populated areas, the density distributions are quite close to each other in both size groups for liberals and industrials, respectively. Hence, from this point of view, no significant systematic differences are to be expected in the following IV regressions between the defined agglomeration groups. Note that this fact cannot be inferred from the summary statistics in table B.1 because the table shows weighted values.

Figure 5 shows the dispersion of AMCs across all four agglomeration groups as well as the the development of the AMCs' population over time. Each of the three scatter plots shows the size of the AMCs in either 1872 or 1920 on the axes. The vertical and horizontal straight lines denote the median of the variable on the respective axis, which is exactly the value that classifies AMCs as *Large* or *Small*. The drawn 45° line serves as an orientation as to whether and how much a region has grown. All of the three graphs indicate that some AMCs with below average population, have grown substantially and climbed into the higher size class, while the contrary is also visible. These transitions are unsurprisingly more frequent the larger the considered period is. Certainly, the size of regions is highly persistent, however the transitions of AMCs between the different groups confirm that there is room for explanations as to why a region has taken either a good or an unfavorable growth path.

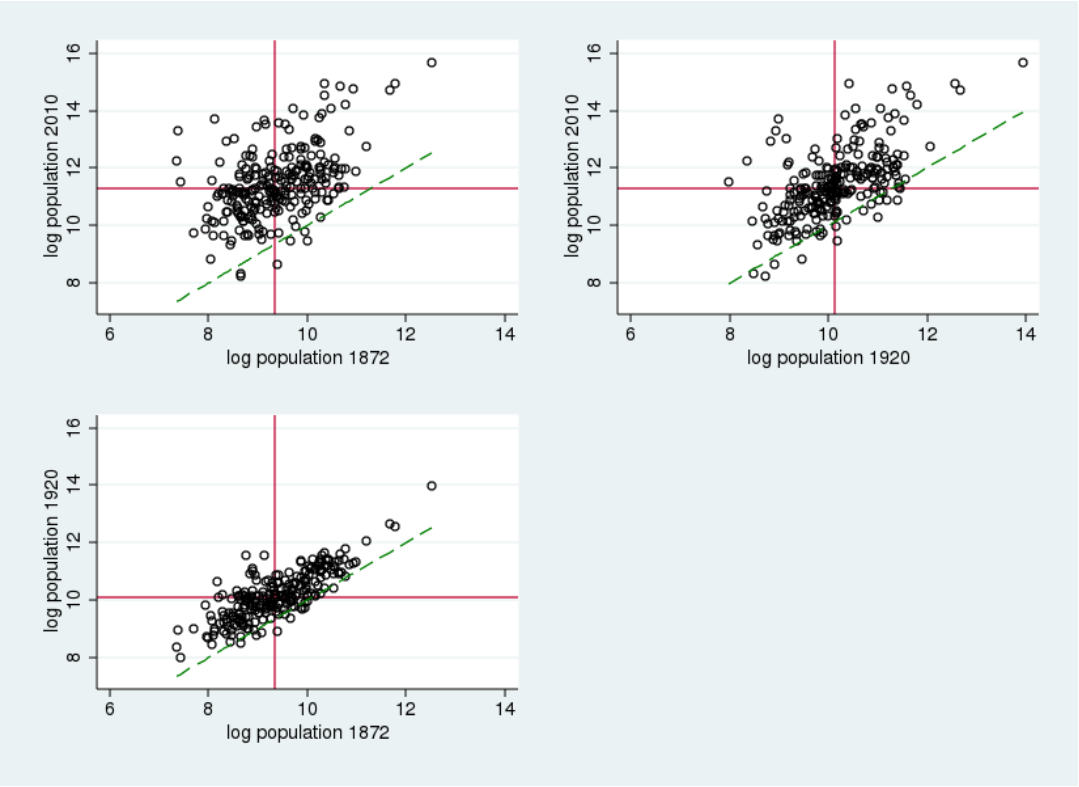
5.4.2 IV results for AMC size groups

Table 3 present the estimations after the division of the sample according to the agglomeration groups defined above. We report OLS and IV estimates for the face-to-face skill concentration. In the first two columns, we split the sample according to the AMC's size in 1872 and consequentially only apply both instruments from 1872. Analogously, columns (3) to (4) refer to one of the 1920's size groups and include the two IVs from 1920 instead. Each estimation contains the full set of control variables, including the 1^{st} nature advantage controls. Because the number of clusters is close to the number of regressors (72) in these GMM estimations, we had to partial out some variables (the UF dummies) in order to obtain weighting matrices of full rank. By the Frisch-Waugh-Lovell theorem, the remaining coefficients are unchanged, however (Baum *et al.* 2007). Yet Stock-Yogo weak IV statistics cannot be calculated and we resort to the simple 1. stage F-value.

Three patterns stand out. First, no matter which of the two group definitions is used, in each IV estimation, only the skill concentration for the largest group shows a significant coefficient at the 1% level. Second, these coefficients are comparable to the previous IV estimations. In fact, the coefficient from the estimation with the 1872 IVs is somewhat lower than in table 2 but, as we have seen previously, this coefficient is estimated with less precision than those in the IV regressions with the IVs from 1920.

The first pattern is partially explained by the fact that the OLS coefficients in the largest agglomeration group are also the most significant. Another OLS coefficient is significant

Figure 5: Development of AMCs across size groups



Notes: The circles in each graph represent the AMCs' population in different years. The vertical and horizontal straight lines denote the median of the variable on the respective axis, i.e., the value that divides the AMCs in the sample into the group of *Large* or *Small* AMCs. The dashed lines represent a 45° line through the origin.

at the 1% level which, however, cannot be confirmed in the IV estimation. This observation suggests that agglomeration economies arise only in highly populated regions and, conversely, externalities are not significantly measurable in regions with sparse population. So it seems that there is a kind of 'critical mass' for the occurrence of these externalities.

The third pattern concerns the relevance of the instruments. Only in the group of the most populated AMCs is there a substantial relation between the concentration of face-to-face skills and the supply of manufacturing and liberal professions in the previous century. Note that in table 3 both R^2 and the first-stage F-values are close to zero for the *Small* AMCs. In contrast, these first-stage statistics in the group of the *Large* AMCs are satisfactory and even higher than in the full sample. Only the K statistic indicates paradoxically that the 1872 IVs are weak, even though the first stage F-value amounts to 23. These facts also explain why a significant externality is identified only for this group. As before, the results for the analytic skill concentrations are relegated to the appendix (table B.3) but follow the described pattern of the face-to-face skill concentration.

To sum up, note first that the relevance of the IVs is independent of the existence of agglomeration economies and hence can be interpreted separately. The fact that the instruments only have significance in one of the groups is a clear indication that an interaction between the population size of a region and its economic development path exists. This observa-

Table 3: IV regressions – agglomeration groups – face-to-face skills

	Dependent variable: log wage p.h.			
	(1)	(2)	(3)	(4)
group:	Large in 1872	Small in 1872	Large in 1920	Small in 1920
	OLS			
face-to-face conc.	0.628***	0.447***	0.632***	0.172
	(0.100)	(0.156)	(0.090)	(0.127)
R^2	0.374	0.376	0.372	0.375
	1872 IVs		1920 IVs	
face-to-face conc.	0.653***	0.850	1.027***	1.506
	(0.130)	(1.956)	(0.088)	(5.792)
	1.-stage statistics			
1. R^2-part.	0.351	0.006	0.516	0.003
1. F-stat	23.280	0.242	68.930	0.095
K	1.577	0.259	6.550	0.135
K-p	0.209	0.611	0.011	0.713
Hanson J	2.318	0.140	0.857	0.587
Hanson J-p	0.128	0.708	0.355	0.444
Obs.	898,139	392,305	1,000,010	293,036

Notes: The sample is split according to the median AMC population size in 1872 and 1920 in columns (1), (2) and (3), (4), respectively. All regressions include the basic control variables specified in table 2 and the 1^{st} nature advantage proxies. Regressions are weighted by the Census population weights. UF dummies are partialed out due to the high number of exogenous variables relative to the number of clusters. Standard errors in brackets are clustered at the AMC-level. * denotes significance at ten, ** at five and *** at one percent level.

tion provides an indication about the mechanism by means of which industrial and liberal professions have transformed into modern industries over time. Interaction between supply (factories) and consumer demand (population) created a virtuous cycle for the local economy. In a wider perspective our findings suggest that market potential and history do play a crucial role in the development of agglomerations, as argued by theories of the New Economic Geography.

5.5 Heterogeneous effects

As a last extension, we analyze whether the agglomeration effects are heterogeneous in different segments of the population. So far we have seen that workers generate more or less pronounced spillovers in accordance with their professional activity. Likewise, workers may also differ as recipients of these spillovers.

Duranton and Puga (2001) and Glaeser (1999) argue that experimentation and learning are fundamental in the generation of agglomeration economies and consequentially one might expect that workers with better education would benefit to a greater extent. We

distinguish workers depending on their level of schooling in the following way. The Brazilian educational system offers a division into five different levels of schooling, which we used previously to generate dummy variables in the estimations. We run our preferred specification separately for each of these five education groups. It is important to note that the skill concentration variable remains unchanged in this exercise. That means, we do not want to measure how workers from different education groups interact with and benefit from their local peers. Instead, we distinguish how much different workers benefit from their factual environment.

Table 4 contains the main results for the interpersonal skill concentrations. All estimated agglomeration economies' coefficients are significant at the 1% level and the first-stage F-value of at least 30 shows that the IVs are highly relevant in all education groups. It is also striking that the wage externality is highest in the group of workers with the highest education, i.e. university graduates. This result confirms the hypothesis that people in this group are more apt to absorb knowledge quickly and to use it to their own advantage.

Table 4: IV regressions – face-to-face skills – by worker's schooling

	Dependent variable: log wage p.h.				
	(1)	(2)	(3)	(4)	(5)
schooling:	< 4 **years**	$4-7$ **years**	$8-10$ **years**	$11-14$ **years**	$15+$ **years**
face-to-face conc.	0.985***	0.702***	0.788***	1.001***	1.991***
	(0.136)	(0.093)	(0.095)	(0.074)	(0.184)
1.stage F-stat.	31.610	40.850	51.640	56.990	46.890
Obs.	241,789	248,327	255,001	436,900	108,427

Notes: The sample is split according to the educational attainment of workers. All regressions include the basic control variables specified in table 2 (except education) as well as the 1^{st} nature advantage proxies. Regressions are weighted by the Census population weights. Occupation fixed effects are partialed out due to the high number of exogenous variables relative to the number of clusters. Standard errors in brackets are clustered at the AMC-level. * denotes significance at ten, ** at five and *** at one percent level.

Furthermore, a polarization of wage effects is visible. Besides the university and college graduates, the group with the next highest wage benefits is the one with the lowest level of education. One possible explanation is that there are strong complementarities between these groups, that is, the low-skilled workers benefit from productivity increases of high-skilled workers and vice versa. The evidence in Eeckhout *et al.* (2014b) also support this kind of complementarity between workers with the highest and lowest skills across US cities. Another possible reason is known as Baumol's cost disease of the personal services. Simple personal services, that require craftsmanship but not necessarily formal schooling, are bound to the general wage level in a city due to rents and transport costs. If rent and wage levels are higher, the prices and wages of these workers also increase, although their productivity is not directly affected by the analytical or interpersonal skills of other workers.

Wage gains in the two middle education groups are lower, but still positive. Yet a slight increase in school duration is still visible. The effects that we identify are static and compatible with a variety of reasons as to why workers may benefit from the concentration of skills and high-skilled occupations. Because the benefits are not only limited to those oc-

cupational groups that generate the externalities, we conclude that matching alone cannot be responsible for all of the benefits, but rather sharing or learning play an important role, too. Again, these observations are visible in case of both the concentration of analytical skills and the concentration of high-skill occupations, cf. supplementary tables B.4 and B.5. Thus we are confident that these observations reflect an advantage of education that can be generalized.

5.6 Robustness

5.6.1 The concentration of high-skilled occupations

As a first and extensive robustness check in this section, we substitute the concentration of skills for the concentration of certain occupations. To be accurate, our measure of occupational concentration is defined as the log of the number of people in a certain occupation per 1000 workers within each region. As explained previously on the basis of figure A.1 in section 4.3, those occupation groups with the highest remuneration are appropriate substitutes because they also exhibit the highest scores in face-to-face and analytical skills. These three occupational groups are: managers & directors, scientists and medium-skilled technicians. Supplementary figure B.5 in Appendix B illustrates the spatial distribution of these jobs. Unequivocally, managers, scientists and medium-skilled technicians are also most frequently encountered in heavily populated regions. In this representation, the concentrations of these professions also vary quite substantially – between 1 and 267 per 1,000 inhabitants in an AMC. Continuing the examples in section 5.1.1, the municipality Areias registers 17 managers and 3 skilled technicians per 1,000 inhabitants whereas these statistics are equal to 116 and 86 in Florianópolis.

The information on occupations does not emerge from a worker survey, as skill measures do. In fact, the concentration of occupations is a simple and objective count that is directly derived from official data. Our skill measures may possibly be exposed to criticism simply because we import them from a US workforce survey, although the application of instrumental variable estimators overcomes potential measurement error in the skill scores. The occupation concentration only contains measurement error to the extent that workers' jobs are misclassified despite the utilization of the most disaggregated occupation classification. Only if such erroneous entries do not occur at random across regions – and we have no reason to presume that – does this pose a threat to our identification strategy. In any case the occupation concentration is an endogenous variable – just like population size – and thus instrumental variables are required. Still, in our view, skills are a preferable measure because skills provide a more intuitive understanding as to why wage externalities emerge. Observing that the regional concentration of managers involves higher remuneration does not offer a direct indication for why precisely such wage effects occur and thus no useful policy implications can be derived. However, the individuals and firms may improve their proper skill composition and thus generate and benefit from higher wage externalities, as the present paper demonstrates.

For the sake of space, we move directly to the IV estimations with these occupation concentration variables. Supplementary figures B.6, B.7 and B.8 illustrate the first stage correlations between the four IVs and the concentration of managers, scientists and skilled technicians. This time the explanatory power and the strength of the link is even stronger. For the occupation group of science – which is most directly related in the generation of knowledge – both the R^2 and the coefficient for the effect of a 1% increase in liberal professions in 1920 arrive at a value of 0.66.

Table 5 presents the results from OLS regressions and from our preferred IV estimations when the concentration of occupations are utilized to detect agglomeration economies. Basically, the results outline the same scenario that we saw previously. On the one hand, the exogeneity and relevance of the instruments are confirmed by the test statistics. On the other hand, the occupations confirm that the positive externalities of analytical and interpersonal skills are not driven by the definition or the use of US skill data. The concentration of managers, occupations in science and skilled technical staff has an unambiguously positive effect on local earnings. The level of this wage elasticity with respect to the number of people in these respective occupations varies between 0.15 and 0.29.

In these specifications, measurement errors exist only to the extent to which the occupational classification in the Brazilian Census is inaccurate. Assuming that the latter has no systematic flaws across regions, the difference between the OLS and IV estimates in table 5 are entirely due to unobservable characteristics and the endogeneity bias. Therefore, one can roughly evaluate the portions of the latter biases and measurement error in the previous estimates with the skill variables. Note that the OLS coefficients in table 5 are also lower than those of the IV estimates. This time, the effect increases by between 5% and 30% whereas in table 2 the increase lies between 15% and 64%. The observed increase differential between skill and occupation variables thus goes back to the measurement errors in the skill variables.

This robustness check is also useful because the interpretation of the coefficients is more illustrative than that of skill concentrations. For example, an increase in the proportion of managers in a region by 1% implies average wage gains in the regions working population of almost 0.3%. The positive impact of scientists and skilled technicians is somewhat lower. To make these numbers more concrete, we can say that increasing number of manager of an average AMC by one standard deviation – from 55 to 77 per 1000 inhabitants – is associated with wage increases from 21 to 28 R$ (about 5 and 7 USD) per hour, or a gain of 33%. By means of the concentration of interpersonal skills, we derived from table 2 that an increase of +88 managers per 1000 inhabitants would raise wages by +10%. Thus, the latter calculation is likely to underestimate the positive effect of managers because the estimation draws on the *average* skill value of the population. When the wage impact of managers alone is calculated, the necessary increase in the number of managers for an average wage gain of +10% is much lower.

For completeness, supplementary table B.5 reports the wage effects of the three occupation concentrations where we distinguish workers by their education. Again, the positive effect

is by far the largest for university graduates. Workers in this group benefit at least twice as much from the concentration of managers, scientists or skilled technicians than individuals with lower educational achievements. The polarization of wage effects along the duration of schooling is also visible for all three occupation groups. Workers without a complete primary education generally experience the second largest wage externalities from the coagglomeration of high-skilled workers and thus about as much as workers with a college degree.

Table 5: Basic IV regressions – occupation concentrations

	Dependent variable: log wage p.h.					
	(1)	(2)	(3)	(4)	(5)	(6)
	OLS	all IVs	OLS	all IVs	OLS	all IVs
occupation:	managers		science		technical	
occupation	0.221***	0.289***	0.146***	0.153***	0.221***	0.271***
conc.	(0.018)	(0.031)	(0.011)	(0.013)	(0.015)	(0.028)
	1.-stage statistics					
1. R^2-part.	0.503		0.624		0.396	
weak IV: F	49.550		76.670		33.200	
weak IV: τ=5%	21.530		20.490		20.390	
K	9.380		9.279		18.220	
K-p	0.002		0.002		0.000	
Hanson J	3.247		1.489		6.493	
Hanson J-p	0.355		0.685		0.090	

Notes: The regressions include the basic control variables specified in table 2 as well as 1^{st} nature advantage proxies. Regressions are weighted by the Census population weights. Standard errors in brackets are clustered at the AMC-level. The number of observations varies slightly due to a few missing values of the IV but it lies above 1,290k in all of the estimations. * denotes significance at ten, ** at five and *** at one percent level.

The fact that similar results are obtained throughout this first robustness check indicates clearly that the origin of the observed agglomeration externalities is not tied to one specific occupation group or the provision of certain institutions like headquarters or research institutions but rather to skills that are common across the three occupation groups.

5.6.2 Sectoral variation

For the next robustness check, we exclude the agricultural sector from our analysis. The skill values of the profession 'agricultural producer' are problematic because, according to the US skill classification, these workers primarily require managerial and entrepreneurial skills, e.g., for selling commodities on spot markets, making investments in expensive machinery. Consequentially, large-scale agriculture producers rightfully have the second highest scores in analytical skills and above average values in face-to-face skills. Whether this generous description also applies to small-scale farmers is questionable, however. Figure A.1 shows that agricultural producers are located primarily at the bottom of the wage

distribution but also in its middle suggesting both types are present in the data. Thus, the assignment of high analytical and interactive skills to all agricultural producers may have systematically exaggerated and be the origin of measurement error. On the other hand, the bias should not be too large because the major occupation category in the agriculture sector is of simple agricultural *workers*, who earn little and have unquestionably less analytical and interactive skills. To analyze if or to what extent the identification of agglomeration economies is distorted thereby, all calculations and estimations have been repeated without this sector.

As a related robustness check, we focus exclusively on the manufacturing sector and repeat the entire analysis once again. The manufacturing sector is usually more in the focus of studies about agglomeration economies. Nevertheless we are convinced that other sectors may generate externalities, too, especially when thinking about face-to-face communication skills. If spillovers between different sectors – specifically between the manufacturing sector and the remaining branches of the economy – indeed existed and the calculation of the skill concentrations were now exclusively based on the employees in only a single sector, this sample reduction would then introduce an omitted variable bias and the estimates should differ from the previous ones.

In fact, the former robustness check is contained within the latter since the relevant sample is merely reduced further. The first two columns in table 6 show the effects of the concentration of face-to-face skills from IV estimations in both reduced samples. In both columns the agglomeration economies of the skill concentrations are still positive and significant. Again, the Hanson J statistic indicates that our instruments are exogenous. The first-stage F-value is still sufficiently high in column (1). In the manufacturing sample the F-value is below 10 but since the skill concentration coefficient is highly significant and almost identical to the one in column (1), we see our previous estimates reconfirmed.

In comparison to the previous estimate in the last column of table 2, the effect of interactive skills is registered as lower by 0.19. Nevertheless, we are more interested in the economic significance of the effect than in the mere size of these coefficients. Note that due to the modified sample size, the skill concentrations were re-calculated and re-centered on their mean in the new sample. The standard deviation of the face-to-face skill concentration in this new sample is equal to 0.098. Hence, the effect of an increase by 1 standard deviation evaluated at the sample mean is equal to 11% and thus the results are very close to those in the baseline case (10%). On the one hand, those observations suggests that the positive externalities are comparable in different sectors of the economy and on the other hand, the robustness checks reassure us of the stability of the statistical relationships.

5.6.3 Alternative explanations

Another concern regarding our identification strategy may be that we are merely capturing a size effect of the region. We have shown that analytic and face-to-face skills are largely concentrated in agglomerations. In general, the relative population size of regions is also quite stable over time. Besides, headquarters, universities and other research institutions

Table 6: Robustness – IV regressions – skill concentrations

	Dependent variable: log wage p.h.			
	(1)	(2)	(3)	(4)
IVs:	all IVs	all IVs	all IVs + 1872 size	all IVs + 1920 size
sample:	no agric.	only manuf.	full sample	
face-to-face conc.	1.188***	1.156***	0.887**	0.747*
	(0.120)	(0.194)	(0.396)	(0.385)
log size			0.013	0.023
			(0.027)	(0.026)
	1.-stage statistics			
1. R^2-part.	0.495	0.243	0.515	0.508
1. F-stat	46.553	8.134	47.500	52.570
2. R^2-part.			0.643	0.609
2. F-stat			50.110	47.780
K	1.959	1.662	29.070	30.390
K-p	0.162	0.197	0.000	0.000
Hanson J	3.536	5.861	4.523	4.207
Hanson J-p	0.316	0.119	0.210	0.240
Obs.	1,145,677	218,792	1,277,941	1,290,444

Notes: The regressions include the basic control variables specified in table 2, 1^{st} nature advantage proxies and are weighted by the Census population weights. The four historical IVs are used in all estimations. In column (3) and (4) we additionally instrument for the AMCs' log population size in 2010 by using either log AMC population in 1872 or in 1920. In column (1) the agriculture sector is excluded from the sample and in column (2) the estimation is restricted to the manufacturing sector. Standard errors in brackets are clustered at the AMC-level. * denotes significance at ten, ** at five and *** at one percent level.

are traditionally located in large cities. Therefore, it could be that the development of modern industries was not directly affected by the concentration of manufacturing and liberal professions, but rather indirectly by the diversity and the innovative environment in agglomerations. To tackle this concern, we provide two exercises.

First, we additionally control and instrument for the population size of the region. As noted previously, population size is a frequently used proxy for several observably equivalent agglomeration externalities (Duranton and Puga 2004). At the same time, this exercise reveals whether the positive effect of skill concentration is independent from such general agglomeration economies. Obviously, both the size and the skill concentration are endogenous variables in our wage regression and thus both require an adequate instrumental variable. IV estimations with more than one endogenous variable are rarely encountered in empirical economics. They can be hard to interpret and are not our preferred choice. Nevertheless, estimation is technically possible, we have sufficient relevant instruments and we are thus able to perform this robustness check.

Columns (3) and (4) in table 6 show the results from the estimations where the AMC population size in the year 2010 is instrumented with its values in either 1872 or in 1920, together with the four historical trades IVs. It turns out that the population size does not significantly affect individual wages once the interpersonal skill concentration and individual characteristics are controlled for. This occurs despite the fact that the instruments for the population size in 2010 are relevant according to the F-value in this second first-stage

estimation (2. F-stat.). The remainder first-stage statistics also demonstrate that this extended set of instruments is relevant and exogenous. More important is that the concentration of interpersonal skills still exert a significantly positive wage externality over and above the most common proxy for agglomeration economies: population size. Although the additional control reduced the skill concentration effect, it is close to its original value of 1.

For the second robustness check in this subsection, we use the historical size of AMCs as an additional instrument for the current interpersonal skill concentration to test whether size has in fact been more important than the concentration of historic trades. If our previous four instruments merely reflect a size effect of the region, then one would expect that the relevance of the IVs would now rise. Due to the high correlation between the population sizes in 1872 and 1920, we restrict the set of instruments to those from 1920 because they have shown higher explanatory power thus far. To conclude, we show the corresponding results for all of the five skill and occupation concentration measures in table 7. Although significant wage effects are obtained in each case, the population size definitely does not dominate the concentration of historical trades in 1920 in the first-stage regressions. Population size is a significant predictor at the 1% significance level in two of the five cases but the concentration of liberal professions is about at least as important. Moreover, the F-value in these five different first-stage regressions does not differ markedly from those in the previous estimations. These observations suggest that the historical size does not contribute much further explanatory power to today's concentration of skills and occupations beyond the one of historical trades. All in all, the size and significance of the wage externalities of interpersonal skills proved stable throughout these robustness checks and neither the historical size of regions, nor definitions of skills, areas, industrial composition, etc. altered the main findings in this paper.

6 Conclusion

The present paper provides the first causal evidence that the concentration of interpersonal and analytic skills generate positive wage externalities. Our estimates suggest that an increase of one standard deviation in the concentration of these skills would raise the average wage level in a region by about 10%. These agglomeration economies especially accrue to university graduates pointing out that high-skilled workers are also most capable of capitalizing the local spillovers. Nevertheless, positive wage effects are measurable for workers in all education groups. It thus seems reasonable that learning and information sharing are crucial for the transmission of these externalities. Further research on the details of the transmission mechanism is certainly required.

To overcome the common problems of measurement error in skills and the endogeneity of the region's skill concentrations, we construct a new set of historical instruments. The data show that the regions with a high density of liberal professions (professors, lawyers, etc.) and industrial occupations over 90 years ago, on average, nowadays host a larger

Table 7: IV regressions – population size as an additional IV

	Dependent variable: log wage p.h.				
	(1)	(2)	(3)	(4)	(5)
skill/occ.:	face-to-face	analytical	managers	science	techn.
skill/occ.	0.983***	1.247***	0.284***	0.154***	0.281***
	(0.092)	(0.163)	(0.030)	(0.013)	(0.027)
	1.-stage statistics				
1920: log of	0.027**	0.010	0.074*	0.189***	0.223***
ind. per 1000	(0.013)	(0.013)	(0.040)	(0.067)	(0.058)
1920: log of	0.061***	0.032**	0.193***	0.364***	0.154**
liberal per 1000	(0.014)	(0.015)	(0.051)	(0.088)	(0.061)
1920: log of	0.008	0.027***	0.062***	0.072*	−0.019
population	(0.005)	(0.006)	(0.022)	(0.039)	(0.028)
1. R^2-part.	0.455	0.451	0.519	0.631	0.393
1. F-stat	77.030	34.830	52.270	79.430	31.430
K	24.230	0.163	17.430	13.260	19.700
K-p	0.000	0.686	0.000	0.000	0.000
Hanson J	0.812	10.750	3.375	2.158	5.938
Hanson J-p	0.666	0.005	0.185	0.340	0.051

Notes: The regressions include the control variables specified in tables 2, 1^st nature controls and are weighted by the Census population weights. Standard errors in brackets are clustered at the AMC-level. The number of observations is equal to 1,293,046 in all estimations. * denotes significance at ten, ** at five and *** at one percent level.

proportion of high-skilled workers which primarily performing analytical and interpersonal skills. This development path particularly takes place in regions that already had a large population in the past. Our data is thus highly suggestive that linkages between localized supply of knowledge and physical production have promoted sustainable economic growth and continue to shape the occupational and earnings structure of regions up to the present day. These channels of long-run development have thus far received less attention compared to the role of institutions. Since the data stem from a single country, formal differences in the legal system and the rule of law between regions were and still are absent. Yet in light of the importance of lawyers and judges in the past it is likely that some regional disparities in the de facto implementation of institutions existed.

References

Acemoglu, D. and Autor, D. H. (2011). Skills, tasks and technologies: Implications for employment and earnings. In O. Ashenfelter and D. Card (eds.), *Handbook of Labor Economics*, vol. 4, Elsevier – North Holland, pp. 1043–1171.

—, Johnson, S. and Robinson, J. A. (2001). The colonial origins of comparative development: An empirical investigation. *American Economic Review*, **91** (5), 1369–1401.

Andersson, M., Klaesson, J. and Larsson, J. P. (2014). The sources of the urban

wage premium by worker skills: Spatial sorting or agglomeration economies? *Papers in Regional Science*, **93** (4), 727–747.

Arrow, K. J. (1962). The economic implications of learning by doing. *Review of Economic Studies*, **29** (3), 155–173.

Autor, D. H., Katz, L. F. and Kearney, M. S. (2008). Trends in US wage inequality: Revising the revisionists. *Review of Economics and Statistics*, **90** (2), 300–323.

—, Levy, F. and Murnane, R. J. (2003). The skill content of recent technological change: An empirical investigation. *Quarterly Journal of Economics*, **118** (4), 1279–1333.

Bacolod, M., Blum, B. S. and Strange, W. C. (2009a). Skills in the city. *Journal of Urban Economics*, **65** (2), 136–153.

—, — and — (2009b). Urban interactions: soft skills versus specialization. *Journal of Economic Geography*, **9** (2), 227–262.

Baum, C. F., Schaffer, M. E. and Stillman, S. (2007). Enhanced routines for instrumental variables/GMM estimation and testing. *Stata Journal*, **7** (4), 465–506.

Botelho, T. R. (2005). Censos e construção nacional no Brasil Imperial. *Tempo Social*, **17** (1), 321–341.

Chaudhary, L., Musacchio, A., Nafziger, S. and Yan, S. (2012). Big BRICs, weak foundations: The beginning of public elementary education in Brazil, Russia, India, and China. *Explorations in Economic History*, **49** (2), 221–240.

Ciccone, A. and Hall, R. E. (1996). Productivity and the density of economic activity. *American Economic Review*, **86** (1), 54–70.

Combes, P., Duranton, G. and Gobillon, L. (2008). Spatial wage disparities: Sorting matters! *Journal of Urban Economics*, **63** (2), 723–742.

Combes, P. P., Duranton, G., Gobillon, L. and Roux, S. (2010). Estimating agglomeration economies with history, geology, and worker effects. In E. L. Glaeser (ed.), *Agglomeration Economics*, University of Chicago Press, pp. 15–66.

Combes, P.-P. and Gobillon, L. (2015). The empirics of agglomeration economies. In G. Duranton, V. Henderson and W. Strange (eds.), *Handbook of Urban and Regional Economics*, vol. 5, Elsevier – North Holland.

Dell, M. (2010). The persistent effects of Peru's mining Mita. *Econometrica*, **78** (6), 1863–1903.

Duranton, G. and Puga, D. (2001). Nursery cities: Urban diversity, process innovation, and the life cycle of products. *American Economic Review*, pp. 1454–1477.

— and — (2004). Micro-foundations of urban agglomeration economies. In V. Henderson and J. F. Thisse (eds.), *Handbook of Regional and Urban Economics*, vol. 4, Elsevier – North Holland, pp. 2063–2117.

Eeckhout, J., Pinheiro, R. and Schmidheiny, K. (2014a). Spatial sorting. *Journal of Political Economy*, **122** (3), 554–620.

—, — and — (2014b). Spatial sorting. *Journal of Political Economy*, **122** (3), 554–620.

Ehrl, P. (2014a). A breakdown of residual wage inequality in Germany. *BGPE Discussion Paper*, **150**.

— (2014b). Task trade and the employment pattern: the offshoring and onshoring of brazilian firms. *BGPE Discussion Paper*, **151**.

— (2015). Minimum Comparable Areas for the period 1872–2010: An aggregation of Brazilian municipalities. *mimeo*.

Fávero, M. d. L. d. A. (2006). A universidade no Brasil: das origens à reforma universitária de 1968. *Educar*, **28**, 17–36.

Finlay, K., Magnusson, L. and Schaffer, M. (2013). weakiv: Weakinstrument-robust tests and confidence intervals for instrumental-variable (IV) estimation of linear, probit and tobit models. http://ideas. repec. org/c/boc/bocode/s457684.html.

Firpo, S., Fortin, N. M. and Lemieux, T. (2011). Occupational tasks and changes in the wage structure. *IZA Discussion Paper No. 5542*.

Florida, R., Mellander, C., Stolarick, K. and Ross, A. (2012). Cities, skills and wages. *Journal of Economic Geography*, **12** (2), 355–377.

Fujita, M., Krugman, P. R. and Venables, A. J. (2001). *The spatial economy: Cities, regions, and international trade*. MIT press.

— and Thisse, J.-F. (2002). *Economics of Agglomeration: Cities, Industrial Location, and Regional Growth*. Cambridge University Press.

Glaeser, E. (1999). Learning in cities. *Journal of Urban Economics*, **46** (2), 254–277.

— and Maré, D. (2001). Cities and skills. *Journal of Labor Economics*, **19** (2), 316–342.

Glaeser, E. L., Kallal, H. D., Scheinkman, J. A. and Shleifer, A. (1992). Growth in cities. *Journal of Political Economy*, **100** (6), 1126–1152.

Henderson, V., Kuncoro, A. and Turner, M. (1995). Industrial development in cities. *Journal of Political Economy*, **103** (5), 1067–1090.

IBGE (2010). *Censo Demográfico 2010 – Notas Metodológicas*. IBGE: Rio de Janeiro.

Kang, T. H. (2010). *Instituições, voz política e atraso educacional no Brasil, 1930-1964*. Ph.D. thesis, Universidade de São Paulo.

Levy, M. S. F. (1974). O papel da migração internacional na evolução da população brasileira (1872 a 1972). *Revista de Saúde Pública*, **8** (supl.), 49–90.

Maciente, A. N. (2013). *The determinants of agglomeration in Brazil*. Ph.D. thesis, University of Illinois at Urbana-Champaign.

Marcondes, R. L. (2012). O mercado brasileiro do século xix: uma visão por meio do comércio de cabotagem. *Revista de Economia Política*, **32** (1), 142–166.

Marshall, A. (1890). *Principles of Economics*. London: Macmillan, 1961, 9th edn.

Mello, P. C. d. (1984). *A economia da escravidão nas fazendas de café: 1850-1888*. PNPE: Rio de Janeiro.

Michaels, G., Rauch, F. and Redding, S. J. (2013). Task specialization in US cities from 1880-2000. *NBER Working Paper*, **18715**.

Monasterio, L. M. and Reis, E. (2008). Mudanças na concentração espacial das ocupações nas atividades manufatureiras no Brasil: 1872-1920. *IPEA Texto para discussão*, **1361**.

Moulton, B. R. (1986). Random group effects and the precision of regression estimates. *Journal of Econometrics*, **32** (3), 385–397.

Naritomi, J., Soares, R. R. and Assunção, J. J. (2012). Institutional development and colonial heritage within brazil. *Journal of Economic History*, **72** (02), 393–422.

Nunn, N. (2008). Slavery, inequality, and economic development in the Americas: An examination of the Engerman-Sokoloff hypothesis. In E. Helpman (ed.), *Institutions and Economic Performance*, Harvard University Press.

Olea, J. L. M. and Pflüger, C. (2013). A robust test for weak instruments. *Journal of Business & Economic Statistics*, **31** (3), 358–369.

Reis, E. (2014). Spatial income inequality in Brazil, 1872–2000. *Economia*, **15** (2), 119–140.

—, Cossio, F. B., Morandi, L., Medina, M. and Abreu, M. (2002). *O Século XX nas contas nacionais*. IBGE: Brasília.

—, Pimentel, M., Alvarenga, A. I. and dos Santos M. C. H. (2011). Áreas mínimas comparáveis para os períodos intercensitários de 1872 a 2000. In *1º Simpósio Brasileiro de Cartografia histórica*.

Rosenthal, S. S. and Strange, W. C. (2004). Evidence on the nature and sources of agglomeration economies. In V. Henderson and J. F. Thisse (eds.), *Handbook of Regional and Urban Economics*, vol. 4, Elsevier – North Holland, pp. 2119–2171.

— and — (2008). The attenuation of human capital spillovers. *Journal of Urban Economics*, **64** (2), 373–389.

Spitz-Oener, A. (2006). Technical change, job tasks, and rising educational demands: Looking outside the wage structure. *Journal of Labor Economics*, **24** (2), 235–270.

Stock, J. H. and Yogo, M. (2005). Testing for weak instruments in linear IV regression. In D. W. Andrews and J. H. Stock (eds.), *Identification and inference for econometric models: Essays in honor of Thomas Rothenberg*, Cambridge University Press, Cambridge, pp. 80–108.

Stolz, Y., Baten, J. and Botelho, T. (2013). Growth effects of nineteenth-century mass migrations: "Fome Zero" for Brazil? *European Review of Economic History*, **17** (1), 95–121.

The Maddison-Project (2010). Statistics on world population, GDP and per capita GDP, 1-2008 AD.

von Thünen, J. H. (1826/1966). *Der isolierte Staat in Beziehung auf Landwirtschaft und Nationalökonomie*. Perthes – Hamburg, translated from German by C. M. Wartenberg (1966) Von Thünen's Isolated State, Oxford – Pergamon Press.

Appendices

A Further details on skill measures

Unfortunately, there is no representative workforce survey in Brazil that would provide information about the skills workers require in their occupations. In the US there are two such surveys that ask workers to state the importance of various ability requirements and activities performed in their professional life. The DOT is used by Autor *et al.* (2003) and Acemoglu and Autor (2011) use its predecessor, the O*NET, to replicate their skill definitions. Maciente (2013) provides a matching of US occupations to Brazilian occupations at the most disaggregated (6-digit) level using their names and synonyms. As a further validation, each occupation code from both countries is compared to the international ISCO88 classification, for both of which a transition already exists. Despite the efforts, Maciente (2013) admits that the mapping of occupations may introduce unpredictable measurement error in the skill scores. By far and large this will not change that fact that a truck driver requires more manual and less analytical skills than a bank accountant which in turn requires less interactive and analytical skills than his director.

We initially distinguished between analytical, cognitive and manual skills based on the definitions provided in Acemoglu and Autor (2011). For example, the score of analytical skills is composed of the importance of the following elements in the O*NET survey: "analyzing data or information", "thinking creatively" and "interpreting the meaning of information for others". Additionally, we follow Firpo *et al.* (2011) and introduce "face-to-face" skills. Some elements in the definitions of cognitive and interactive skill overlap. Both definitions include "establishing and maintaining interpersonal relationships" and "coaching and developing others". Additionally, the cognitive skill measure captures "guiding, directing, and motivating subordinates" and "how important being very exact or highly accurate" is. Face-to-face skills in turn include "assisting and caring for others", "working directly with the public", as well as the frequency of "face-to-face discussions" The skill values in each occupation are calculated as the mean of the aforementioned O*NET elements and finally standardized to have a mean of 10 and a standard deviation of 1 in the occupation data.

To avoid repetitive descriptions and results, the study focuses only on two and, over the further course of the study, only on one type of skill measure. As noted above, there is some overlap between cognitive and face-to-face skills. Aside, manual skills are largely mirror-image to those other skills. A large intensity of manual skills may as well be interpreted as the absence of analytic and interpersonal skills. The correlation between manual skills and analytic and face-to-face skills is equal to -0.59 and -0.72, respectively. The focus and presentation of two different skills is sufficiently informative, the more so as the other two skill measures do not add different information. Cognitive skills basically reproduce the results of face-to-face skills and manual skills yield a reverse image.[15] For the presentation

[15] Manual skills are mainly concentrated in rural areas and the coefficient of their concentration in wage

of the results, we chose the two types of skills with the largest degree of differentiation, namely face-to-face and analytical skills.

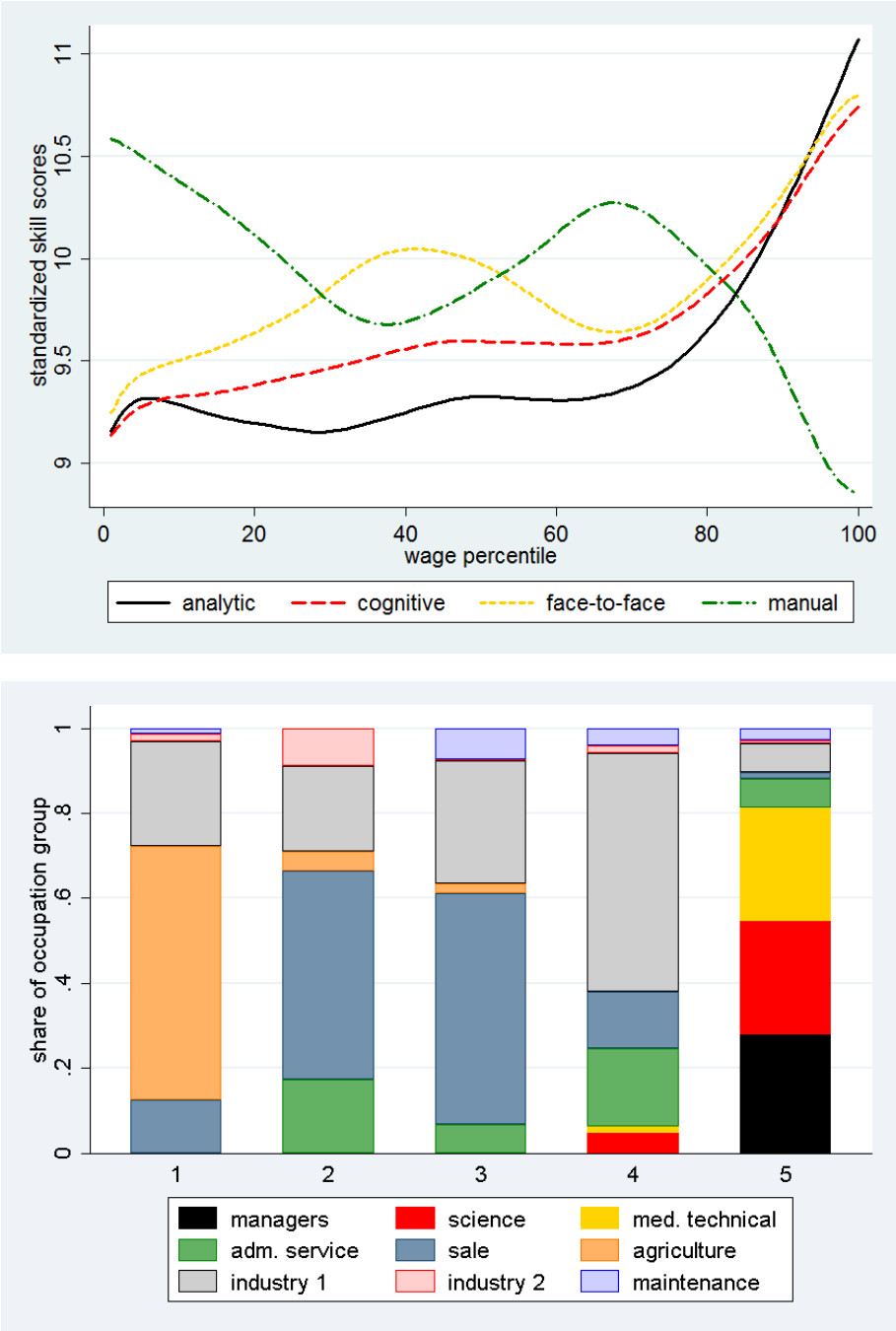
The upper plot in figure A.1 shows the average skill scores along the population weighted wage distribution in 2010. The lower graph is useful to make sense of the distribution of skills above. It shows the employment shares of 1-digit occupation groups along the same wage distribution aggregated to quintiles. It is noteworthy that analytic and cognitive skill measures are almost parallel. As expected, workers in the bottom fifth of the wage distribution score the highest values in manual skills and exhibit low values for all other skills. These workers are mainly engaged in agriculture and in simple industrial activities. Coherently, at the upper end of the wage distribution, the highest values of analytic, cognitive and face-to-face skills are to be found. Mainly managers, researchers and people with a medium-skilled technical occupation are located in the top fifth of the wage distribution. Workers further to the left have clearly lower and almost monotonically falling values of analytic and cognitive skills.

Between the 2^{nd} and 3^{rd} quintiles, face-to-face skills are predominant, what comes from the fact that many sales jobs are located there. Many occupations in the manufacturing industry obtain higher earnings. The latter demand high manual skills but also require somewhat more analytical and cognitive skills than those professions with lower income in sales and agriculture, for example. Naturally, the workers in the industry rely less on interactive / face-to-face skills. All in all, the picture reflects the Brazilian peculiarities, such as a large agricultural sector, but it seems reasonable and in line with both the expectations and findings from, for example, Germany (cf. Spitz-Oener (2006)) or the US (cf. Autor *et al.* (2008)).

For completeness, we present the spatial concentration of manual and cognitive skills, as well as the correlation between the local wage level and the skill indices analog to section 5.1. Figure A.2 reveals that large AMCs have the lowest average values for manual skills, or vice versa, physical skills are unsurprisingly most utilized in rural regions. Cognitive skills, in contrast, are concentrated in largely populated regions, just like analytical and interactive skills. Figure A.3 confirms that there is a significant and astonishingly linear relation between the log wage and both manual and cognitive skills.

regressions is negative. This negative coefficient is in line with (Bacolod *et al.* 2009a: 145): "motor skills have a hedonic price that decreases with MSA population". Even though motor skills and manual skills are meant to capture the same, the variables in Bacolod *et al.* (2009a) are defined differently because another, anterior data base (DOT) provides the importance of skills in professions. That is, the more workers that use manual skills intensively are concentrated in one place, the lower the wage. One explanation is that employers gain more bargain power over the workers, since manual skills are largely unskilled activities. Moreover, a large supply of low-skilled workers might depress those wages substantially.

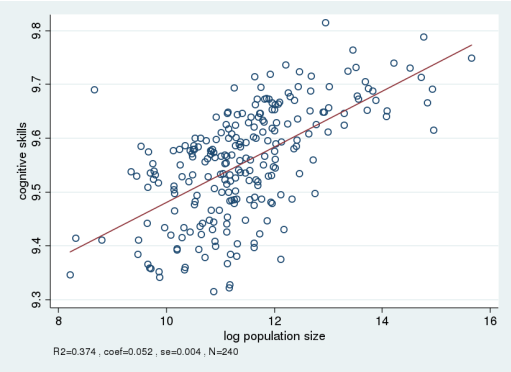
Figure A.1: Skills and occupations along the wage distribution



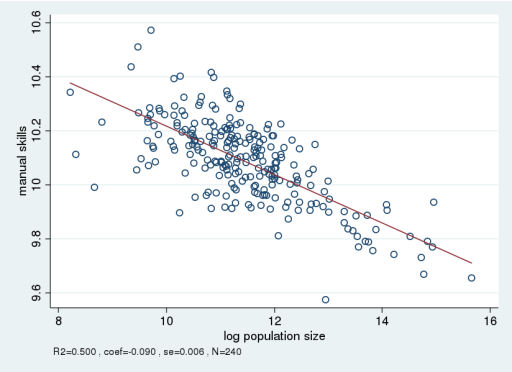
Notes: The upper graph shows the standardized skill scores based on the O*NET along the wage distribution for the entire working population interviewed for the Brazilian Census in 2010 using a kernel weighted regression. The lower graph aggregates the wage percentiles to quintiles and shows the share of workers in each occupation group (1-digit classification).

Figure A.2: Spatial concentration of skills (2) – AMC means

(a) cognitive



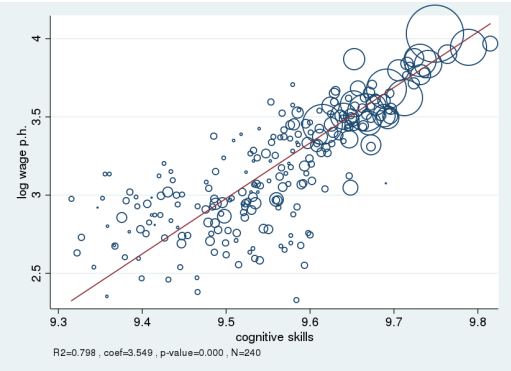
(b) manual



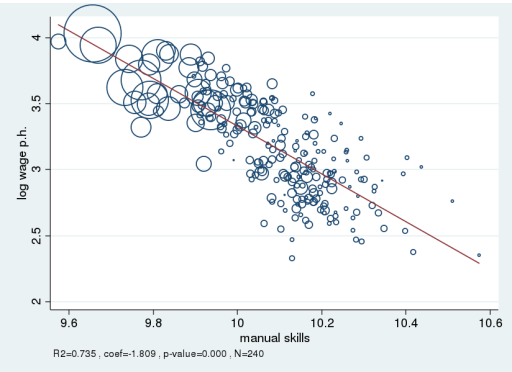
Notes: The circles in each graph represent the AMCs' cognitive and manual skill averages and their log population size for all AMCs with an area smaller than 2,500 sq km. The results from the corresponding (unweighted) linear regression are indicated below each graph.

Figure A.3: Correlation between wages and concentration of skills (2) – AMC means

(a) cognitive



(b) manual



Notes: The circles in each graph represent the AMCs' cognitive and manual skill and log wage averages using population weights. Only AMCs with an area smaller than 2,500 sq km are part of the sample. The results from the corresponding (unweighted) linear regression are reported below each graph.

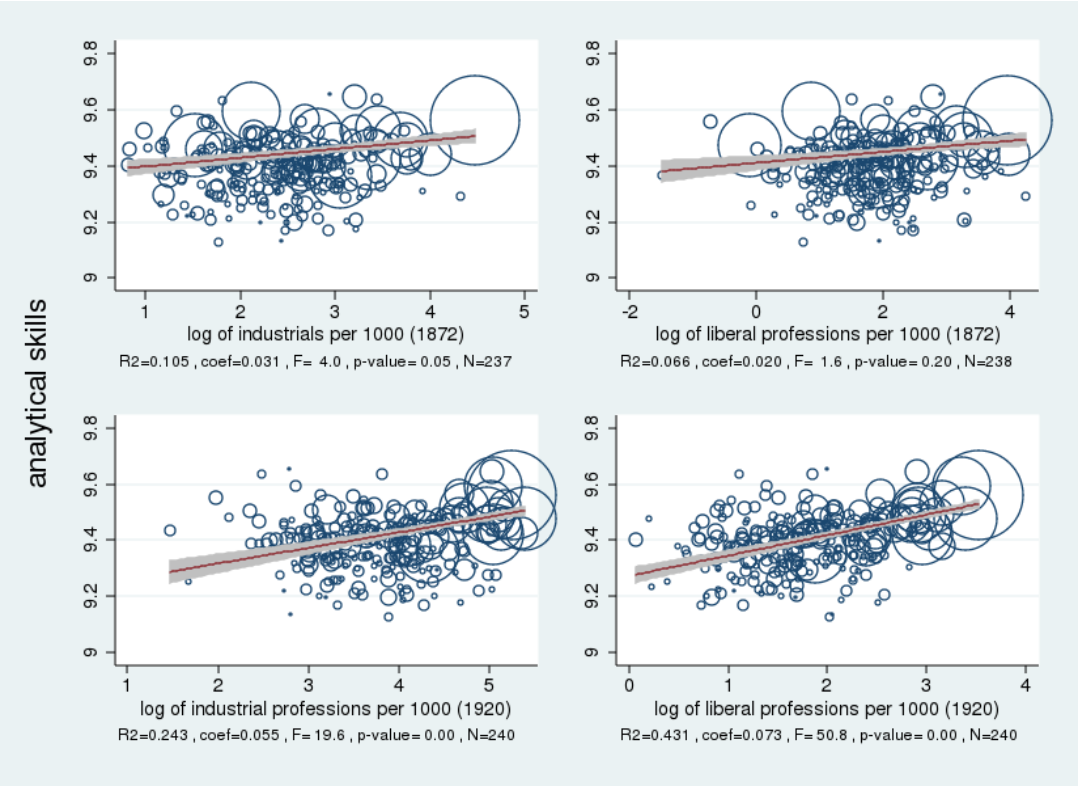
B Supplementary tables and figures

Table B.1: Summary statistics

	(1) Total	(2) Large in 1872	(3) Small in 1872	(4) Large in 1920	(5) Small in 1920
log wage p.h.	3.054	3.057	3.046	3.077	2.940
	(0.924)	(0.939)	(0.875)	(0.927)	(0.904)
analytical skills	9.460	9.469	9.429	9.470	9.409
	(0.645)	(0.647)	(0.635)	(0.648)	(0.626)
face-to-face skills	9.891	9.911	9.828	9.906	9.817
	(0.584)	(0.578)	(0.600)	(0.581)	(0.593)
analytical conc.	9.459	9.468	9.428	9.469	9.408
	(0.085)	(0.084)	(0.080)	(0.083)	(0.076)
face-to-face conc.	9.925	9.944	9.861	9.940	9.851
	(0.105)	(0.099)	(0.098)	(0.102)	(0.083)
managers conc.	4.011	4.031	3.946	4.060	3.766
	(0.338)	(0.330)	(0.357)	(0.304)	(0.392)
science conc.	4.157	4.221	3.942	4.242	3.728
	(0.602)	(0.592)	(0.583)	(0.561)	(0.615)
techn. conc.	4.061	4.113	3.889	4.123	3.748
	(0.431)	(0.391)	(0.506)	(0.376)	(0.538)
log of industrialists 1872	2.829	2.941	2.460	2.923	2.354
	(0.897)	(0.944)	(0.582)	(0.921)	(0.554)
log of liberals 1872	2.214	2.285	1.980	2.293	1.813
	(1.122)	(1.187)	(0.834)	(1.158)	(0.812)
log of industrial occ. 1920	4.483	4.631	3.994	4.631	3.742
	(0.761)	(0.709)	(0.722)	(0.689)	(0.667)
log of liberals 1920	2.532	2.629	2.213	2.661	1.885
	(0.750)	(0.758)	(0.624)	(0.706)	(0.619)
log size 1872	10.274	10.757	8.693	10.583	8.694
	(1.232)	(0.953)	(0.507)	(1.068)	(0.690)
log population 1920	11.194	11.565	9.970	11.547	9.429
	(1.331)	(1.220)	(0.871)	(1.155)	(0.480)
log population 2010	13.580	13.972	12.285	13.910	11.927
	(1.474)	(1.351)	(1.069)	(1.315)	(1.059)
Obs.	1.293m	0.899m	0.393m	1.000m	0.293m
Obs. weighted	15.476m	11.880m	3.596m	12.902m	2.573m

Notes: The table presents the weighted sample mean and the standard deviation (in parenthesis below) for the main variables used in the present analysis. The last two rows indicate the number of individuals in the sample and how many individuals (both in millions) these observations represent when the population weights are applied. The statistics in the first column refer to the working population living in AMCs with an area of less than 2,500 sq km based on Census data from 2010, as defined in sections 4.1 and 4.2. This sample is divided according to the median of the population in 1872 in columns (2) and (3). Columns (4) and (5) refer to a division of the sample according to the median population in 1920, cf. section 5.4.1.

Figure B.1: Correlation between analytical skill mean and historical trades



Notes: The circles in each graph represent the AMCs' analytical skill average and each one of the four instrumental variables in the sample. The results from a weighted linear regression of each of the IVs on the analytical skill concentration are indicated below each graph.

Table B.2: Basic IV regressions – analytical skill concentration

	Dependent variable: log wage p.h.								
	OLS	IV (1)	IV (2)	IV (3)	IV (4)	1872 IVs	1920 IVs	all IVs	all IVs
analytical	0.628***	1.040***	1.149***	1.531***	1.436***	1.100***	1.509***	1.513***	1.380***
skill conc.	(0.079)	(0.263)	(0.284)	(0.209)	(0.212)	(0.254)	(0.193)	(0.186)	(0.186)
	1.-stage statistics								
1872: log of		0.045***				0.028**		0.005	0.004
ind. per 1000		(0.013)				(0.012)		(0.010)	(0.010)
1872: log of			0.035***			0.018		0.008	0.012
liberal per 1000			(0.011)			(0.012)		(0.010)	(0.010)
1920: log of				0.061***			0.002	0.000	−0.002
ind. per 1000				(0.009)			(0.015)	(0.015)	(0.015)
1920: log of					0.069***		0.067***	0.062***	0.064***
liberal per 1000					(0.008)		(0.018)	(0.015)	(0.015)
1. R^2-part.		0.146	0.139	0.295	0.398	0.161	0.398	0.411	0.400
weak IV: F		16.370	13.720	58.350	106.400	11.870	54.080	30.690	28.640
weak IV: τ=5%		37.420	37.420	37.420	37.420	19.300	11.860	21.420	22.170
weak IV: τ=10%		23.110	23.110	23.110	23.110	12.470	8.084	12.980	13.410
weak IV: τ=20%		15.060	15.060	15.060	15.060	8.552	5.896	8.350	8.606
weak IV: τ=30%		12.040	12.040	12.040	12.040	7.062	5.063	6.655	6.847
K						0.818	0.255	0.048	0.093
K-p						0.366	0.613	0.826	0.760
Hanson J						0.370	0.682	2.513	4.678
Hanson J-p						0.543	0.409	0.473	0.197

Notes: The regressions control for a quadratic in worker's age, dummies for occupation, sector, Federal State, education level, occupational position, race, marital status, and whether or not the person is a foreigner, male, illiterate or has a physical deficiency, as specified in section 4. All regressions are weighted by the Census population weights. The estimation in the last column additionally includes the 1^{st} nature advantage proxies. Standard errors in brackets are clustered at the AMC-level. The number of observations varies slightly due to few missing values of the IV but it lies above 1,290k in all of the estimations. * denotes significance at ten, ** at five and *** at one percent level.

Figure B.2: AMCs distinguished by population size in 1872 and 1920

(a) 1872 (b) 1920

Notes: Both graphs are detail of the Brazilian territory showing all AMCs with an area of less than 2,500 sq km. Thereof, those with a population size above the mean in either 1872 (left graph) or 1920 (right graph) are colored in red. AMCs with a size below the respective mean are displayed in green.
Source of the GIS delineation: IBGE.

Figure B.3: Distribution of 1872s IVs across size groups

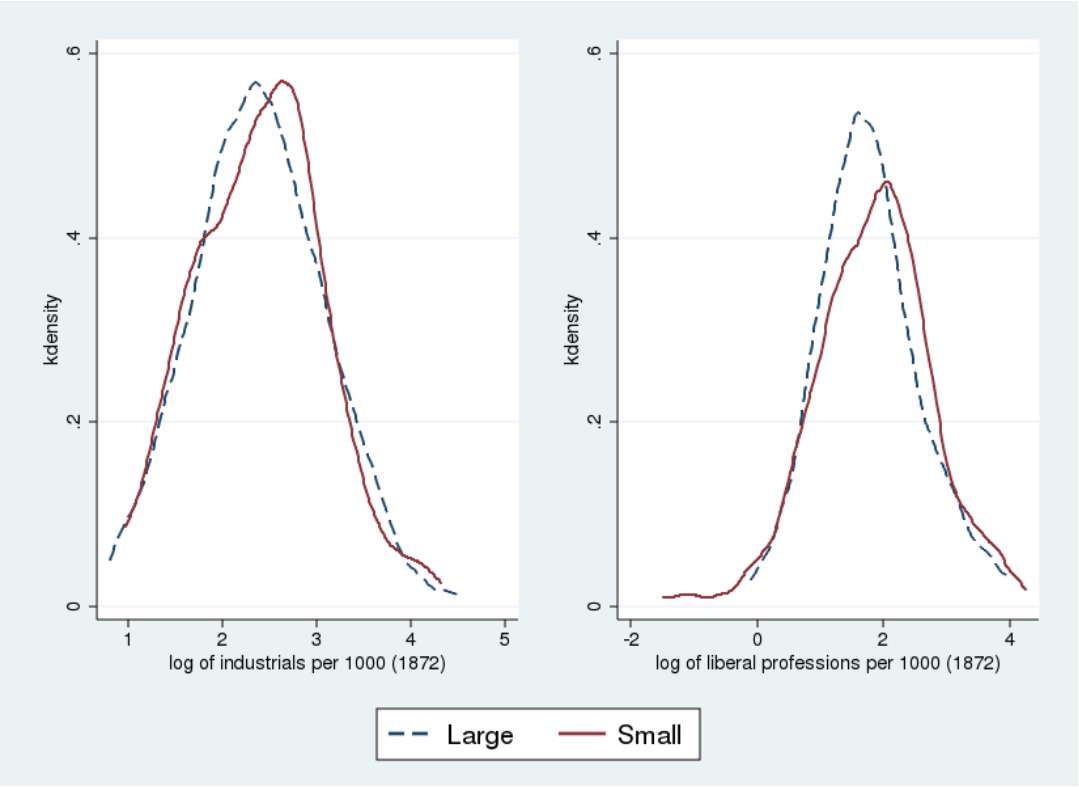


Figure B.4: Distribution of 1920s IVs across size groups

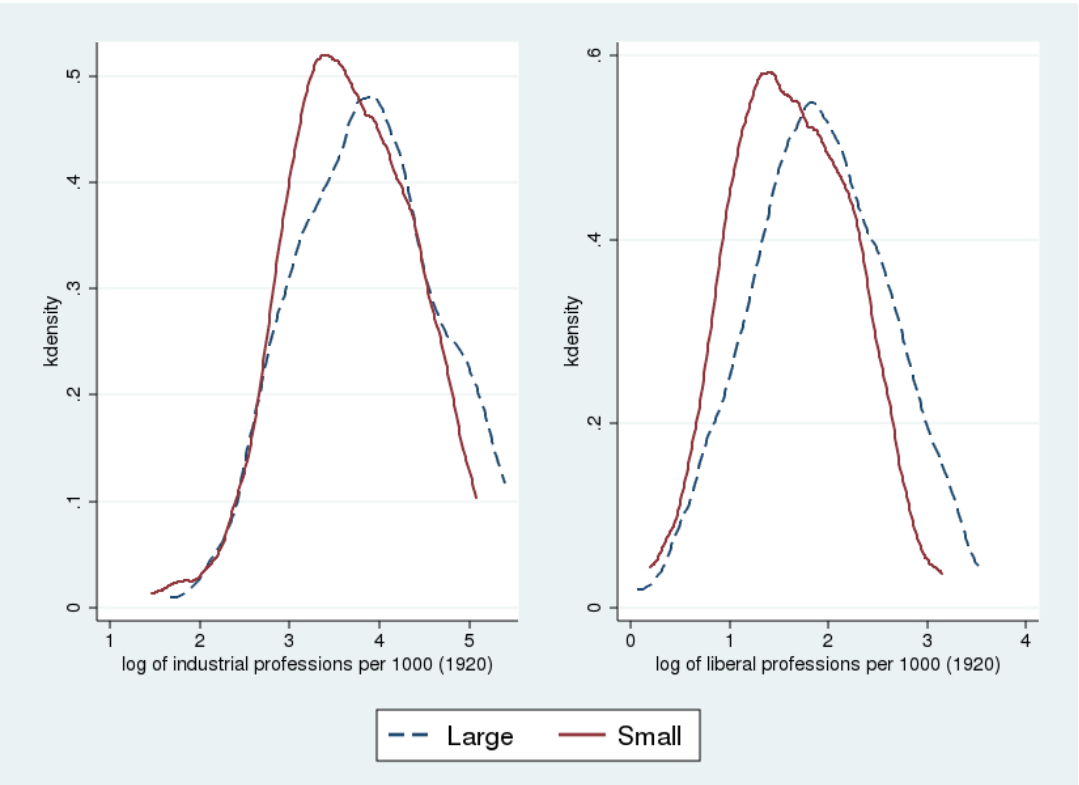


Table B.3: IV regressions – agglomeration groups – analytic skills

	Dependent variable: log wage p.h.			
	(1)	(2)	(3)	(4)
group:	Large in 1872	Small in 1872	Large in 1920	Small in 1920
	OLS			
analytical conc.	0.636***	0.357**	0.712***	0.148
	(0.088)	(0.171)	(0.088)	(0.174)
R^2	0.373	0.375	0.372	0.375
	1872 IVs		1920 IVs	
analytical conc.	0.791***	1.760	1.393***	−0.944
	(0.170)	(4.100)	(0.195)	(1.251)
	1.-stage statistics			
1. R^2-part.	0.355	0.003	0.483	0.043
1. F-stat	12.180	0.082	27.800	2.630
K	0.662	0.419	0.000	0.401
K-p	0.416	0.518	0.986	0.527
Hanson J	0.369	0.001	0.486	0.376
Hanson J-p	0.544	0.975	0.486	0.540
Obs.	898,139	392,305	1,000,010	293,036

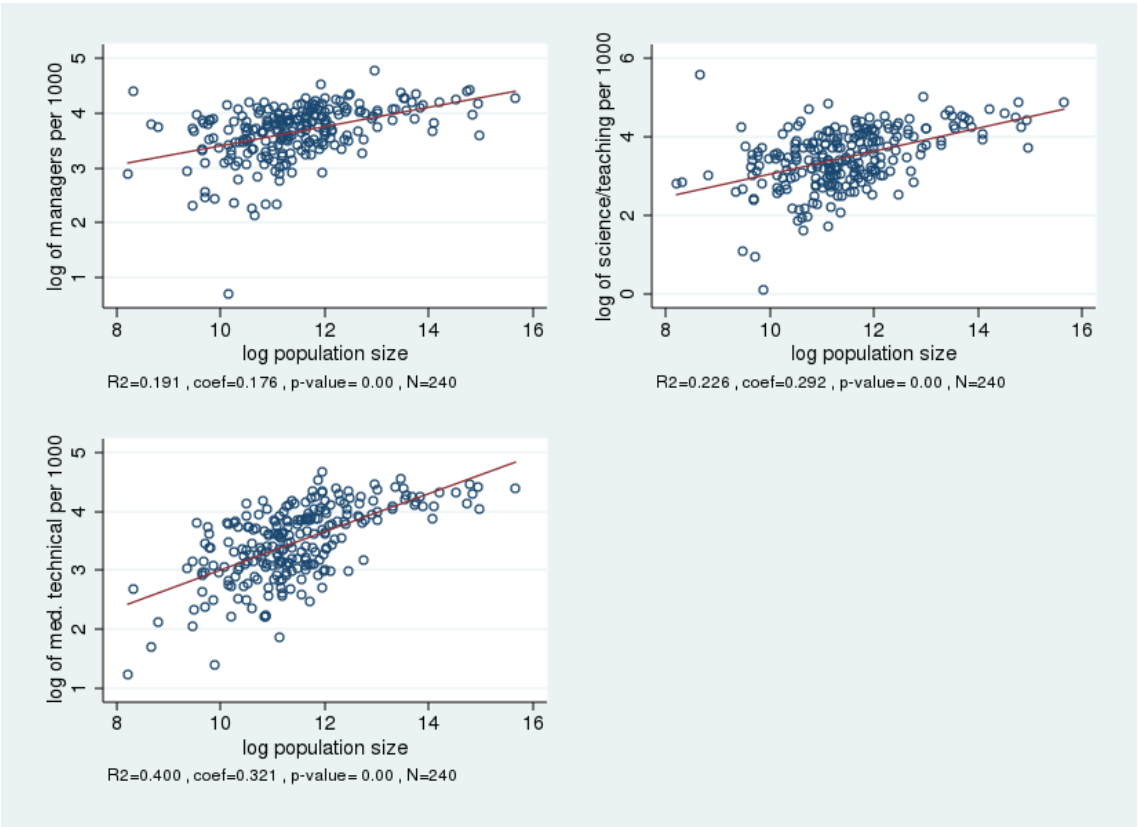
Notes: The sample is split according to the median AMC population size in 1872 and 1920 in columns (1), (2) and (3), (4), respectively. All regressions include the basic control variables specified in table 2 and the 1^{st} nature advantage proxies. Regressions are weighted by the Census population weights. UF dummies are partialed out due to the high number of exogenous variables relative to the number of clusters. Standard errors in brackets are clustered at the AMC-level. * denotes significance at ten, ** at five and *** at one percent level.

Table B.4: IV regressions – analytical skills – by worker's schooling

	Dependent variable: log wage p.h.				
	(1)	(2)	(3)	(4)	(5)
schooling:	< 4 **years**	$4-7$ **years**	$8-10$ **years**	$11-14$ **years**	$15+$ **years**
analytical conc.	1.412***	0.976***	1.096***	1.293***	2.482***
	(0.267)	(0.176)	(0.186)	(0.151)	(0.267)
1. F-stat.	15.610	18.220	16.380	16.540	9.857
Obs.	242,895	248,966	255,379	437,343	108,463

Notes: The sample is split according to the educational attainment of workers. All regressions include the basic control variables specified in table 2 (except education) as well as the 1^{st} nature advantage proxies. Regressions are weighted by the Census population weights. Occupation fixed effects are partialed out due to the high number of exogenous variables relative to the number of clusters. Standard errors in brackets are clustered at the AMC-level. * denotes significance at ten, ** at five and *** at one percent level.

Figure B.5: Spatial concentration of occupations – AMC means



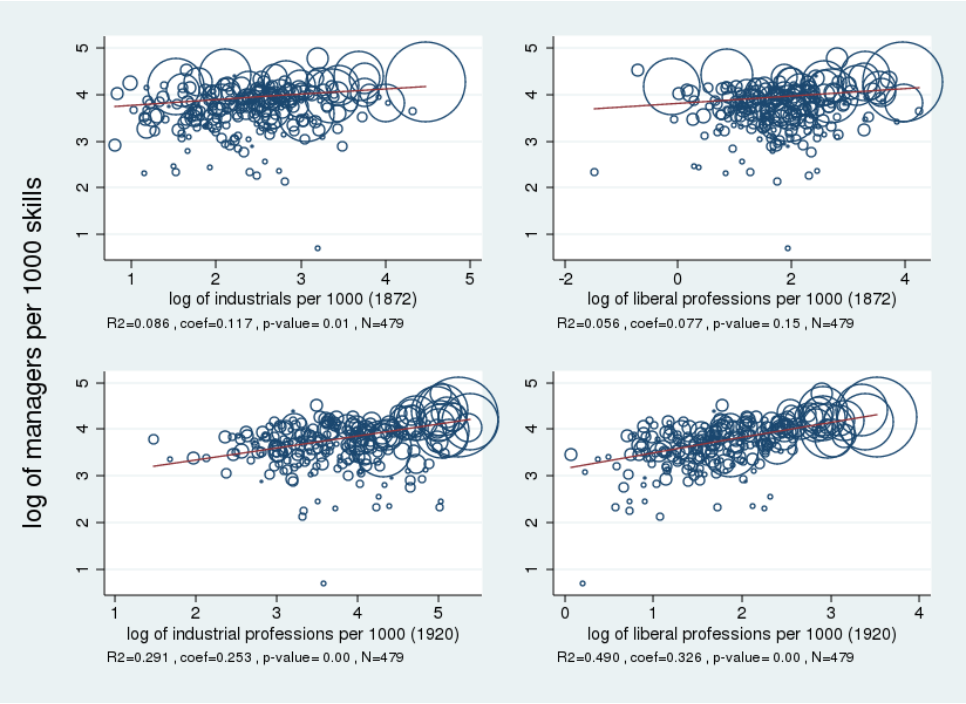
Notes: The circles in each graph represent the AMCs' concentration in each of the three highest skilled occupation (1-digit) groups. Only AMCs with an area smaller than 2,500 sq km are part of the sample. The results from the corresponding (unweighted) linear regression are reported below each graph.

Table B.5: IV regressions – occupations – by worker's schooling

	Dependent variable: log wage p.h.				
	(1)	(2)	(3)	(4)	(5)
schooling:	< 4 **years**	$4-7$ **years**	$8-10$ **years**	$11-14$ **years**	$15+$ **years**
manager conc.	0.269***	0.200***	0.230***	0.283***	0.662***
	(0.040)	(0.030)	(0.033)	(0.026)	(0.069)
1. F-stat	38.220	41.250	38.040	34.610	17.940
science conc.	0.150***	0.106***	0.120***	0.153***	0.343***
	(0.018)	(0.014)	(0.015)	(0.010)	(0.022)
1. F-stat	59.120	63.170	54.990	51.110	24.020
techn. conc.	0.220***	0.185***	0.224***	0.260***	0.597***
	(0.024)	(0.022)	(0.024)	(0.031)	(0.088)
1. F-stat	29.840	25.240	22.970	20.520	9.614
Obs.	241,789	248,327	255,001	436,900	108,427

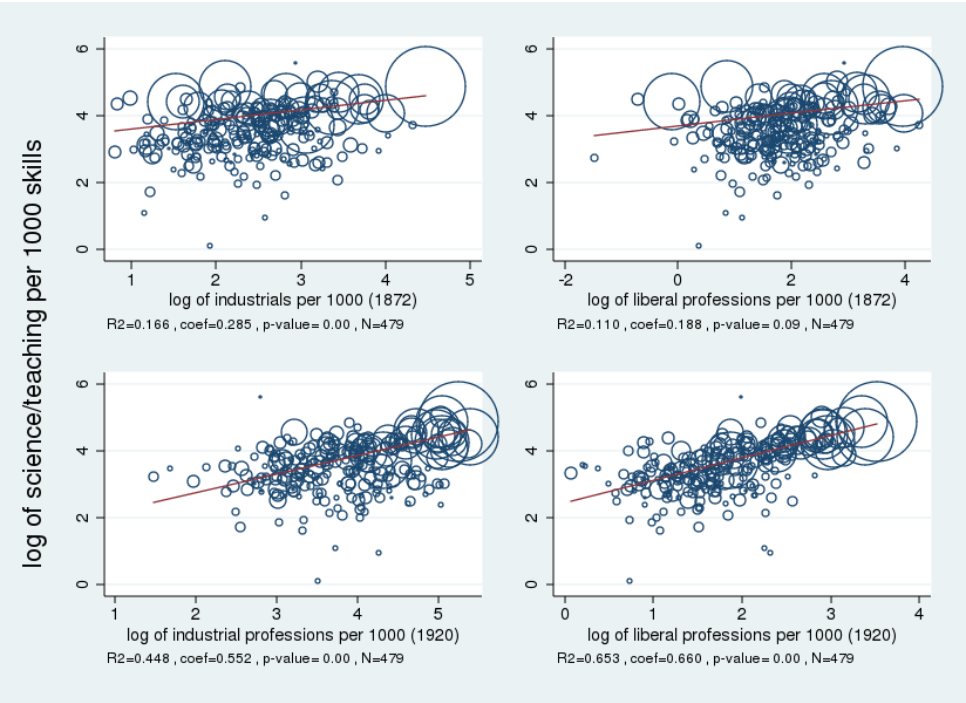
Notes: The sample is split according to the educational attainment of workers. All regressions include the basic control variables specified in table 2 (except for education) as well as the 1^{st} nature advantage proxies. Regressions are weighted by the Census population weights. Occupation fixed effects are partialed out due to the high number of exogenous variables relative to the number of clusters.

Figure B.6: Correlation between share of managers and historical trades



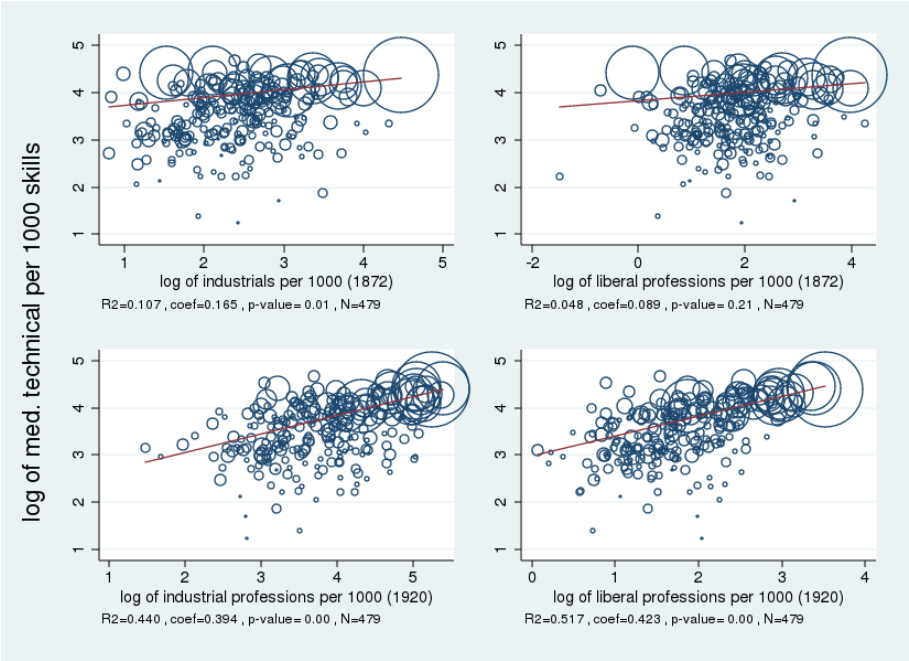
Notes: The circles in each graph represent the AMCs' concentration of managers and each one of the four instrumental variables in the sample. The results from a weighted linear regression of each of the IVs on the concentration of managers are indicated below each graph.

Figure B.7: Correlation between share of scientists and historical trades



Notes: Description is analog to figure B.6 above.

Figure B.8: Correlation between share of skilled technicians and historical trades



Notes: Description is analog to figure B.6 above.